CONTROL-SYSTEM DYNAMICS

WALTER R. EVANS

Systems Group Leader
Electromechanical Engineering Department
North American Aviation, Inc.

McGRAW-HILL BOOK COMPANY, INC.
New York Toronto London
1954

184

Engin. Library

CONTROL-SYSTEM DYNAMICS

Copyright, 1954, by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

Library of Congress Catalog Card Number 53-9877

THE MAPLE PRESS COMPANY, YORK, PA.

Engin. Lib.
Ulrich
10.12.54
87663

PREFACE

The analysis of a given control system is mathematically identical to the analysis of a mechanical shock mount or an electric circuit. Each system can be described by a set of linear differential equations. These equations can be solved by a direct routine for the steady-state response to a sinusoidal input or for the transient response to a step input. The control-system problem differs, however, in that only part of the system is given; the rest of the system must be selected so that the desired response is obtained. This reverse process requires a clear understanding of the problem and many special calculation tricks to be efficient. The purpose of this book is to present these calculation procedures, including the root-locus method developed by the author, to permit the reader to select the best method for his problem. Moreover, the hope is that some of the ideas will help the engineer whose problem is sufficiently complicated that any solution by any method would be welcome.

A book on this subject may be expected to serve both as a textbook for college courses and as a reference text for practicing engineers who want to learn this new field. The book is therefore written so that the reader can start with any of the first seven chapters, depending upon his previous knowledge. Summary statements are given in italics at the beginning of each article to permit more precise selection of topics. This simple option for the reader to omit topics gave the author greater freedom to include physical explanations of concepts which many readers would prefer to accept simply as mathematical facts. Thus Chap. 3 on fundamentals goes all the way back to the reason for the choice of e in expressing a transient as e^{st}. Chapter 4 develops the meaning of transfer functions from close study of the classical solution of simple differential equations rather than by the Laplace transform. Calculations with complex values of s for the root-locus plots in Chap. 7 are then a simple extension of frequency-response concepts. Thévenin's theorem is extended in Chap. 9 to permit isolation of the effect of any system parameter. The conversion from block diagrams to determinants is a gradual one between Chap. 9 and Chap. 10. Finally a proof of Routh's criterion and an explanation of the significance of an autocorrelation function in noise theory are given in the appendixes.

The book contains so many explanations of basic topics, credit is hard to assign. The author is grateful to Profs. Roy S. Glasgow and Frank W.

Bubb, both formerly of Washington University in St. Louis, for their emphasis on understanding rather than superficial knowledge. The General Electric Advanced Engineering Program continued this emphasis, furnished extensive practice in the solution of problems, and provided association with such fellow students as G. E. Walter. At North American Aviation, where the root-locus method was developed, W. D. Mullins gave the initial encouragement by applying the method to autopilot problems immediately. K. R. Jackson, R. M. Osborn, and A. J. Grant provided valuable criticism of the original notes on the subject. J. R. Moore frequently pointed out a need for a more direct approach, which led to many improvements in the method. J. F. Schmidt developed the ideas for a Spirule into the first plastic model. N. F. Parker and D. B. Duncan explained the mathematical theory of noise. C. J. Triska read the manuscript completely and provided valuable criticisms. In justice to some of those mentioned above, the author should point out that this book does not necessarily represent the manner in which they would have presented the same material. Mrs. Frank L. Wiley typed most of the manuscript, a task whose importance any author will appreciate.

Specific references are given where they are known, as to the basic paper by Paul Profos on conformal mapping to find roots. No attempt is made to list the literature on the frequency-response method, because the texts cited do this. Undoubtedly many of the ideas in the chapters on circuit theory and multicoupled systems appear elsewhere, in literature with which the author is not familiar. Apologies are made to the original authors of these ideas, with the explanation that the subjects were investigated independently as required by problems under study.

WALTER R. EVANS

CONTENTS

PREFACE v

TERMINOLOGY xi

CHAPTER 1. THE DYNAMICS PROBLEM

1.1. Basic Form of Control System 2
1.2. Problems Involved in Control Systems 4
1.3. Accuracy Specifications 6
1.4. Fundamentals of Linear Equations 9
1.5. Frequency Response 10
1.6. The Root-locus Method 10
1.7. Nonlinearities 12

CHAPTER 2. DESCRIPTION OF TYPICAL COMPONENTS

2.1. Angle-measuring Devices 13
2.2. Small-motion Pick-offs 15
2.3. Amplifiers 17
2.4. Modulators and Demodulators 19
2.5. D-C Motors 21
2.6. A-C Motors 22
2.7. Solenoids and Clutches 24
2.8. Hydraulic and Pneumatic Components 26
2.9. Choice of Gear Ratios 27
2.10. Summary 29

CHAPTER 3. FUNDAMENTALS OF TRANSIENT RESPONSE

3.1. Mechanical Analogy for a Control System 30
3.2. Time Constant 32
3.3. Exponential Function 33
3.4. Classical Method of Solution 35
3.5. Resonant System 36
3.6. Vector Representation of Sine Waves 38
3.7. Meaning and Use of the Operator $\sqrt{-1}$ 39
3.8. Calculations with Vectors or Complex Numbers 41
3.9. System with Damped Oscillation 43
3.10. Damping Ratio 46
3.11. More Complicated Systems 47

CHAPTER 4. MEANING AND USE OF TRANSFER FUNCTIONS

4.1. Layout of Block Diagram 48
4.2. The Concept of Transfer Functions 50
4.3. Transfer Functions for Exponential Signals 51
4.4. Transient Due to a Unit Impulse 52

4.5. Interchangeability of Blocks . . . 54
4.6. Amplitude of Transient Terms . . . 55
4.7. Exponential Inputs and Initial Conditions . . . 57
4.8. Feedback as Superposition of Signals . . . 59
4.9. Summary . . . 60

CHAPTER 5. THE FREQUENCY-RESPONSE METHOD

5.1. Difficulties with the Classical Method . . . 61
5.2. Frequency-response Concepts . . . 63
5.3. Single-time-delay Components . . . 64
5.4. Quadratic Components . . . 65
5.5. Open-loop Response . . . 67
5.6. Stability . . . 69
5.7. Closed-loop Response . . . 70
5.8. The Effect of a Feedback Function . . . 72
5.9. Logarithmic Plots . . . 73
5.10. Use of Integral Circuit . . . 78
5.11. Use of Lead Circuit . . . 79
5.12. The Question of Transient Response . . . 81

CHAPTER 6. TRANSIENT RESPONSE FROM FREQUENCY RESPONSE

6.1. Asymptotes on Logarithmic Plots . . . 82
6.2. Quadratic Functions . . . 84
6.3. Nyquist Criterion of Stability . . . 85
6.4. Finding Roots by Conformal Mapping . . . 88
6.5. Amplitude of Transient Terms . . . 90
6.6. Complete Solution . . . 92
6.7. Conformal Mapping in Logarithmic Coordinates . . . 94
6.8. An Introduction to the Root-locus Method . . . 95

CHAPTER 7. THE ROOT-LOCUS METHOD

7.1. Roots of the Characteristic Equation . . . 96
7.2. Vector Representation of the Open-loop Function . . . 99
7.3. Sketching the Locus of Roots . . . 100
7.4. Effect of Gain on the Position of the Roots . . . 102
7.5. The Effect of a Lead Term . . . 105
7.6. Aids in Sketching Root-locus Plots . . . 107
7.7. Aids in Calculating Root-locus Plots . . . 108
7.8. The Closed-loop Function . . . 110
7.9. Amplitude of Transient Terms . . . 113
7.10. Use of an Integral Network . . . 115
7.11. Use of Lead Network . . . 117
7.12. Comparison with the Frequency-response Method . . . 119

CHAPTER 8. MULTIPLE-LOOP SYSTEMS

8.1. Tachometer Feedback Loop . . . 122
8.2. Effect of Amplifier Time Constant . . . 124
8.3. Effect of Gain Variations from a Fixed Value . . . 125
8.4. The Fixed Vector Sum of Roots . . . 127
8.5. Classic Problem of Finding Roots of *n*th-degree Equation . . . 129
8.6. The Vibration Absorber . . . 130

8.7. Stabilization of System with Resonant Inner Loop 133
8.8. The Active-compensation Technique 135
8.9. Roll Autopilot Control 137
8.10. Pitch Autopilot Control 139
8.11. The Problem Presented by Multicoupling 142

CHAPTER 9. CIRCUIT ANALYSIS

9.1. Block Diagram of Circuits 146
9.2. Zeros and Poles by Inspection 148
9.3. Use of Thévenin's Theorem 150
9.4. Impedance Characteristics of Passive Networks 153
9.5. Analysis of a Bridged-T Network 156
9.6. Modulated Carrier Signals 158
9.7. Choice of Capacitor in a Complete System 160
9.8. Simultaneous Choice of Two Parameters 163
9.9. Locus versus a Mechanical Parameter 165
9.10. Correlation with Other Methods 166

CHAPTER 10. USE OF DETERMINANTS

10.1. Correlation with Block Diagrams 170
10.2. Meaning of Determinant Manipulations 172
10.3. Development of Cramer's Rule 174
10.4. Statement and Proof of Cramer's Rule 175
10.5. Pitch Autopilot System 178
10.6. Roll-Yaw Coupling 180
10.7. Effect of Small Additional Couplings 182
10.8. The General Problem 183

CHAPTER 11. TRANSFER FUNCTIONS AND TRANSIENT RESPONSE

11.1. Amplitude of Transient Term from Known Zeros and Poles 186
11.2. Transient from Root-locus Plot 188
11.3. The Special Case of Equal Roots 190
11.4. Matching Initial Conditions 192
11.5. Aids in Sketching Transients 193
11.6. Transfer Function from Transient Response 195
11.7. Graphical Solution of Differential Equation 198
11.8. Convolution 199
11.9. Numerical Convolution 201

CHAPTER 12. GRAPHICAL ANALYSIS OF NONLINEAR SYSTEMS

12.1. Sketch of Time Response 205
12.2. Phase-plane Diagrams 207
12.3. Resonant System with Coulomb Friction 208
12.4. Resonant System with Viscous Damping 210
12.5. Relay Time Delay 212
12.6. On-Off Effects in Linear Quadratic Systems 214
12.7. Backlash 215
12.8. Frequency Response of Nonlinear Components 217
12.9. Frequency Response of Nonlinear Systems 221
12.10. Conclusions 222

APPENDIXES

A. The Laplace Transform 225
B. Aids in Sketching Root-locus Plots 230
C. Use of the Spirule 237
D. Routh's Criterion 242
E. Analogy between Flux Plots and Root-locus Plots 247
F. Noise in Terms of Frequency Components 249
G. Autocorrelation Functions 253

PROBLEMS 263

BIBLIOGRAPHY 275

INDEX 279

TERMINOLOGY

Terms are listed below in the order in which the terms arise in the book. Definition usually limits the meaning of a word in the sense used in this book; thus a "differential equation" here is always a linear, ordinary differential equation. A physical picture of transfer functions is presented in this book, and therefore some definitions are different in concept from formal Laplace transform theory but have the same meaning. Typical symbols are given, which are consistent with common usage.[1] Additional comments regarding choice of symbols are made after the following list of terms.

CONTROL SYSTEM

An assembly of devices capable of forcing one quantity to follow the variations of another quantity.

SIGNAL θ, f, v

A variable quantity in a system, such as a voltage, a force, or an angle.

INPUT θ_i, f_i, v_i

The arbitrary signal which is impressed on a system, the independent variable.

OUTPUT θ_o, f_o, v_o

The signal produced by the system. Often it duplicates the input signal as closely as can be achieved.

ERROR θ_e, f_e, v_e

Often the difference between the input and the output signal; in general, the error is the difference between the actual output and the desired output.

FREQUENCY RESPONSE

The characteristic of a system described by the transmission of sinusoidal signals of various frequencies.

TRANSIENT RESPONSE

The natural response of a system due to any disturbance such as a sudden change in the input.

---

[1] In this book the abbreviation "log," rather than "ln," is used for the natural logarithm to the base e. Where any other base is used, it is specified, as in $\log_{10}$.

DIFFERENTIAL EQUATION $\theta_i = \theta_o + \frac{1}{K}\left(1 + T\frac{d}{dt}\right)\frac{d\theta_o}{dt}$

An equation relating signals and their rates of change. In all differential equations treated in this book, the coefficients of all terms are constant.

LINEAR SYSTEM

A system whose characteristics are independent of the amplitude of the signal or of time.

EXPONENTIAL SOLUTION $\theta = \Theta e^{st}$

A universal form of solution for a linear ordinary differential equation, because $de^{st}/dt = se^{st}$.

CHARACTERISTIC EQUATION $0 = 1 + \frac{1}{K}(1 + Ts)s = 1 + \frac{1}{KG(s)}$

An algebraic equation obtained by substitution of a solution of the form e^{st} into the differential equation of the system.

ROOT r_1, r_2

A value of s which satisfies the characteristic equation.

IMAGINARY NUMBER $s = j\omega$

A number which arises in the solution of the equation $s^2 + \omega^2 = 0$.

COMPLEX NUMBER $s = \sigma \pm j\omega$

Numbers which arise in the solution of the equation

$$(s - \sigma)^2 + \omega^2 = 0$$

VECTOR $s = |s|\underline{/\psi}$

A complex number treated in terms of its magnitude $|s|$ and its angle ψ rather than as its real and imaginary components.

BLOCK DIAGRAM

A diagram, used to describe a system, which identifies components in blocks and shows signal paths between blocks. A block should include a logical portion of the system such as motor and load; if these are separated, then shaft torque must be shown as a signal.

TRANSFER FUNCTION $\frac{\Theta_o}{\Theta_e}(s), \frac{\Theta_o}{\Theta_i}(s), \frac{V_2}{V_1}(s)$

The description of a component as the ratio of output to input of that component for the case of an exponential signal e^{st} being transmitted. The transfer function is obtained from the differential equation by substitution of s for the operation d/dt. It may also be used to find the transient response of the component.

UNIT STEP

A sudden change, in a signal, of unit magnitude.

UNIT IMPULSE

A large signal acting for a short time so that the integral of the signal with respect to time is unity.

TRANSFORM $\Theta_o(s)$, $V_e(s)$

The transfer function needed to create a given signal from a unit impulse. This definition includes initial conditions, just as does the formal definition by the Laplace transform integral.

CONFORMAL MAPPING

A plot of a function of s versus a grid of values of the variable s. The Profos method of finding roots is to sketch a curvilinear grid based upon the vector plot of the frequency response.

ROOT LOCUS

For a system whose characteristic equation is $KG(s) + 1 = 0$, the root locus is a plot in the s plane of the values of s which make $G(s)$ a negative real number. A point along the locus is a root if the gain K is selected to make $K|G(s)| = 1$.

ZERO q_1, q_2

A value of s which makes a given function, usually $G(s)$, zero. On root-locus plots, a zero is denoted by a small circle ○.

POLE p_1, p_2

A value of s which makes a given function, usually $G(s)$, infinite. On root-locus plots, a pole is denoted by a cross ×.

MULTILOOP SYSTEM

A system whose block diagram has several loops so that simplification can be achieved by successively replacing a loop with single block.

MULTICOUPLED SYSTEM

A system in which the signal paths on the block diagram are intercoupled to the extent that direct simplification is not possible. Determinants are advantageous for such systems.

DETERMINANT

An array of coefficients of a set of simultaneous equations; the solution can be expressed in terms of the determinant by a convenient set of rules.

CONVOLUTION

The process of finding the total response of a system as the sum of the responses to individual impulses.

NOISE

An unwanted or false input signal.

MEAN-SQUARE VALUE

The average of a square of a variable.

AUTOCORRELATION FUNCTION $\phi(T) = \overline{f(t)f(t + T)}$

The average value of a function multiplied by the same function after it has been offset by a given time, the average value being expressed as a function of this offset time.

The choice of θ_o for output angle and v_e for error voltage allows symbols which are firmly established in various fields to be used with the subscripts to identify the significance of the variables in a control-system sense. The transfer-function symbol $(\Theta_o/V_e)(s)$ identifies both variables exactly and is convenient in multiplication of transfer functions; thus $(\Theta_o/V_e)(s)(V_f/\Theta_o)(s) = (V_f/V_e)(s)$. The transfer function $(\Theta_o/V_e)(s)$ is equal to the ratio of the transforms of the variables $\Theta_o(s)/V_e(s)$ only for the special case of initial conditions being zero.

For root-locus plots the terms zero and pole refer to the function $G(s)$ in the characteristic equation $KG(s) = -1$; the position of the roots along the locus is then a function of K. Roots are denoted by a solid dot $\cdot$ for specific values of K.

CHAPTER 1

THE DYNAMICS PROBLEM

The purpose of this chapter is to outline in general terms the problem of designing a control system to operate quickly and accurately under many different conditions. A gun-turret control system, for example, must make the turret follow the position specified by a computing gun sight for various target motions. The sight is moved suddenly through a large angle when a new target is located, and the gun turret must swing into this new position as quickly as possible. During the interval of tracking the target, the motion of the sight is very uniform, and the error between the computed gun position and the actual gun position must be very small. This gun position must also be held in space in spite of wind forces on the turret and rolling motion of the airplane or ship on which the turret is mounted.

If the system were completely specified and the problem were merely to determine its response under given conditions, books on the subject would not be necessary; the standard method of setting up differential equations and solving them would be sufficient. Actually, however, only part of the system is given; the part which controls the turret motor is free to be selected. The problem is to select this control element in order to minimize errors.

A seemingly universal solution to this problem would be to amplify the error signal, *i.e.*, an input signal proportional to the difference between actual and computed gun positions, to whatever extent is necessary to produce full motor action with an acceptable error. Unfortunately such a control might cause the turret to oscillate back and forth past the neutral position, with an amplitude limited only by the maximum outputs of the amplifier and motor. Such a system is said to be unstable and is unsatisfactory because of errors, excessive heating, and possible overstressing of parts.

The desired solution, which will permit low errors while avoiding instability, must involve not only the instantaneous present value of error, but also its past history and its predicted future value. This type of solution requires a very clear picture of the original system and of the control elements. Two useful approaches to the solution, the frequency-response method and the root-locus method, are described qualitatively at the end of this chapter, after a closer study of the nature of the problem and the simplifications which are common to both methods.

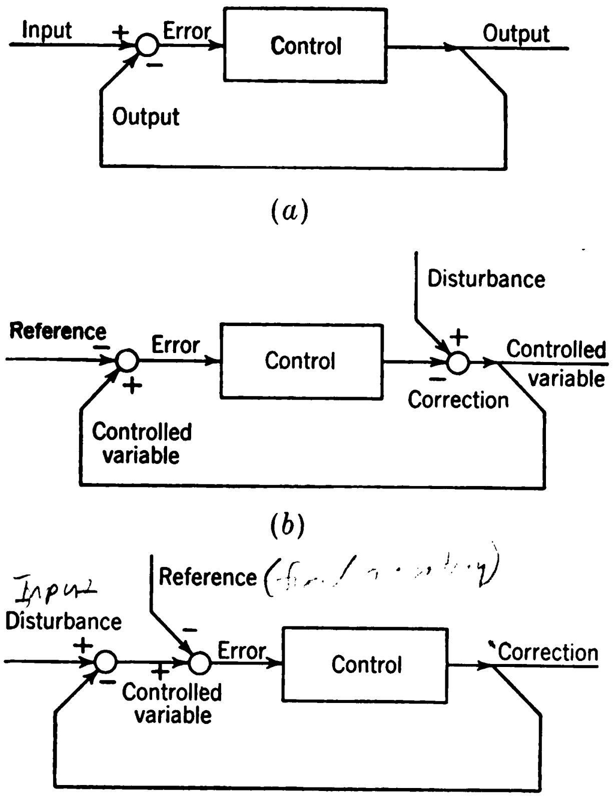
1.1. Basic Form of Control System

> *An automatic control system drives its output primarily as a function of the measured error between an input signal and the actual output. The advantage of the system is that the device providing the input signal is not disturbed, while the output is driven with the power of its actuator and the accuracy of the measuring system.*

A simple gun-turret system is a good example of an automatic control system because the basic elements are clear-cut. The input to the system is the angle of the sight set by the gunner; the output is the angle of the gun turret. The error, *i.e.*, the difference, between angles of sight and turret is measured, amplified, and used to drive the turret into alignment with the sight. The advantage of the system is that the gunner can move the sight easily, while the turret is being driven with the power of its own actuator and with the accuracy of the measuring system. The term "servomechanism" or simply "servo" is used to describe a control system whose function is to make a shaft position follow the variations of a low-power signal.

A regulator is a control system whose principal function is to maintain a quantity constant in spite of various disturbances. A voltage regulator for a d-c generator, for example, compares the terminal voltage of the machine with a fixed reference voltage and makes the difference voltage control the generator field. The servo and the regulator can be shown to involve the same control problem by a comparison of the so-called "block diagrams" in Fig. 1.1. Such diagrams emphasize cause and effect by arrows noting the direction of signal action while denoting parts of the system simply by blocks. For the servo (Fig. 1.1*a*) the difference between the input and output is the error signal to the control block. For the regulator (Fig. 1.1*b*) the difference between the fixed reference and the controlled variable is the error signal to the control block. The output of this block is a correction of the regulator, which, subtracted from the disturbance, gives the value of the controlled variable. The similarity of the regulator and servo can be emphasized by redrawing the block diagram for the regulator (Fig. 1.1*c*). The disturbance is considered to be the input to the system, and the correction is the output; the difference between these two is the controlled variable. The reference is a fixed quantity; therefore the change in the controlled variable from the reference value is the error. A servo acts to make the output follow the input as closely as possible; a regulator acts to make the corrections compensate for the disturbance as closely as possible.

The similarity between servos and regulators is emphasized by control systems which act as both. A gun-turret system, for example, is primarily a servo; yet in so far as the turret position is prevented from being influenced by wind forces, the system is a regulator. An autopilot serves as a regulator in maintaining straight and level flight, but it acts as a servo if the input is changed from a fixed reference to maneuver signals.



(a)

servo

(b)

regulator

(c)

Fig. 1.1. Control systems: (a) servomechanism, (b) regulator, (c) regulator treated as a servomechanism.

The human body is a control system, although its action is so automatic that one scarcely realizes the many functions involved. In picking up an object, the eye detects the error in position between the object and the hand. The signal through the brain to the muscles causes the hand to be moved toward the object until the error is zero. Mechanisms can be used to supplement or completely replace human operations, as will be illustrated by the case of an autopilot. The airplane itself is a supplement to the muscles, in that the positioning of a throttle or a control stick can achieve large changes in velocity. A pilot flying on visual contact with the ground continues to serve both as the error detector and as the brains of the system. Upon entering clouds, however, a gyro provides an artificial horizon as a supplement to the eyes in detecting aircraft attitude; *i.e.*, the gyro assumes part of the error-detecting function. Finally, automatic movement on the control surfaces in response to the gyro horizon may completely replace the pilot by an autopilot for straight and level flight. The pilot is freed from a routine operation to check radio position, engine performance, and flight plan.

The gun-turret system of a bomber is another example of the step-by-step adaptation of mechanisms to perform functions previously performed by a person. The B-17 bomber had handhold guns, with sighting rings and tracer bullets as aids to the gunner in hitting the target. The B-29 replaced the handhold gun with a power turret; the correction for motion of the target was made by a computer; finally radar was added to detect the position of the attacking fighter. Each step of mechanization in this system improved the system: the gunner can do a more accurate job of tracking with an optical sight than with a gun; the computer can make corrections faster and more accurately than the gunner's eye and brain; and the radar can track the target at night or through clouds.

Industrial applications of automatic control systems serve the same functions of eliminating routine and obtaining greater accuracy. The development of any machine involves some form of control. The invention of the steam engine, for example, was followed shortly by the invention of the flyball governor to control its speed. Many machines have self-regulating characteristics to some extent; for instance, a d-c shunt motor tends to run at constant speed regardless of load. Sometimes the need for very accurate speed control may justify expense for the control which is far greater than that of the motor. The production of a paper mill, for example, is limited by the speed which can be maintained by the mill without breaking the continuous strip of paper. One control involves measuring the sag of the paper between rollers by photoelectric cells to control the speed of the next roller.

The applications of control systems are increasing at present for many reasons. The necessary components such as measuring devices, amplifiers, computers, and actuators are available in readily usable form. There are many systems whose speed and accuracy are at present limited by the control. Finally, the knowledge of control systems is becoming more widespread among engineers with the result that more applications are being recognized.

1.2. Problems Involved in Control Systems

The problems involved in control systems cover the full range of invention, analysis, design, and testing. The analysis of the *dynamic response of linear systems is the main theme of this book because it is a single, well-integrated subject which forms a foundation for experience.*

The invention of a control system occasionally calls for insight which the inventor himself may find hard to trace. The flyball governor for speed control of a steam engine, for example, combines measurement of speed, a reference, and control of a valve in one device. Today components are available to permit many control systems to be developed, and new problems continually arise. A gas-turbine drive for a propeller, for example, requires simultaneous control of speed and temperature by adjusting the fuel rate and the pitch of the propeller. In other systems space limitations may require that a device intended to measure the angle of a shaft be used also to apply torques to that shaft. Little is attempted in this book to explain the process of developing a system except to provide a background of examples.

The analysis of many control systems could be routine if the complete system were given and the problem were merely to determine its dynamic response to various input signals. The equations for the system could be set up and combined into a single equation, the solution found, and the result plotted. Inefficiency in the details of the procedure could be tolerated because only one solution would need to be obtained; also the physical picture of the system could be lost in the mathematical manipulations because only the final result would be needed. A typical servo problem, however, has only part of the system specified, with the remainder to be selected so that the error of the system is kept within given limits for given input signals. The choice of a motor in a servo, for example, may be fixed by the top speed and stall torque required; the allowable error for a constant-velocity input is often specified. The simplest control system would simply involve applying across the motor a voltage which is proportional to the error signal. Many systems, however, would develop a violent oscillation; such a system is said to be unstable. The methods of analysis described later in the book explain the cause of the oscillation, suggest the necessary changes in the system, and permit reasonably direct calculation of numerical values. These methods are made the core of the book for reasons summarized at the end of this topic.

The problem of design is one of selection. Design of a control system involves many components, such as measurement devices, amplifiers, and actuators, each of which has characteristic problems and performance. A pair of "synchros," for example, can transmit angle information accurate to $\frac{1}{8}$° even though each synchro is only about 1 in. in diameter. A typical servo motor with a 2-in.-oz stall torque and 10,000 rpm top speed can accelerate unloaded at 40,000 rad/sec^2. The total number of components is so great and the range in their limitations, size, and performance is so wide that the engineer concerned with control cannot be an expert on all of them. Thus this book describes only a few typical components (see Chap. 2), deriving the differential equations needed in later analysis.

Finally, testing a component or a control system for its dynamic response can be accomplished in many ways. For example, the voltage impressed upon a d-c shunt motor may be varied sinusoidally and the corresponding sinusoidal variation in speed noted. The ratio of the peak motor speed attained to the peak voltage applied will be a function of the frequency of variation of the voltage. For a very low frequency of variation, the ratio of peak motor speed to peak applied voltage will be nearly as large as that for fixed applied voltage. For high frequencies, however, the motor will have a limited time to accelerate and the top speed reached will be low. A series of tests at various frequencies will give the "frequency response" of the motor. Another method of testing is to apply a fixed voltage to the motor suddenly and record the build-up of speed. The sudden application of voltage would be said to be a "step input," and the record of build-up of speed would be the "transient response" of the motor. Either test will be shown to define the characteristics of the motor.

This brief summary serves to indicate the great breadth of the problems involved in control systems. The analysis problem is so well adapted for treatment in technical papers and books that one is apt to lose sight of the other problems involved in invention, design, and testing. Fortunately, however, the analysis problem is in many respects the central problem. The analysis, for example, serves to translate the over-all specifications of the system into the individual specifications of the components, thus simplifying design. A searching analysis can often reveal system modifications which will permit easing the specifications of the components. Moreover, a thorough understanding of the various methods of analysis is a great asset in the interpretation of test data. For example, it will be shown that frequency-response data can be interpreted in terms of a transient response, or vice versa, which permits the tests to be conducted in whichever manner is most convenient. For these reasons, the analysis problem is made the central theme of this book.

1.3. Accuracy Specifications

The dynamic accuracy of a system is commonly specified in terms of the error for any one of three forms of input: constant-velocity,

> ***step input, or sinusoidal variations. A specification for one form of input is equivalent to a specification for another form, although the relationship is not simple for systems having three or more points of energy storage.***

The input to a system can be completely arbitrary in form. For convenience in stating specifications, however, three standard forms are selected: constant-velocity (Fig. 1.2*a*), step input (Fig. 1.2*b*), and sinusoidal (Fig. 1.2*c*). The input and output angles θ_i and θ_o are plotted against time t; the error angle θ_e is the difference between input and output angles.[1] The slopes of the curves are, of course, the respective velocities. In Fig. 1.2*a* the input velocity B_i is constant; the output velocity is equal to the input velocity but the output angle lags the input angle by the error angle B_i/K. In Fig. 1.2*b* the input angle is fixed after its initial step, and therefore the error angle decreases as the output angle approaches the input. The output velocity is always proportional to the error angle for this simplified system, however, which determines the transient curve, which is of a type very common in control systems. In Fig. 1.2*c* the output angle is a sine wave which can be shown to lag behind the input-angle sine wave by the angle whose tangent is $2\pi f/K$, where f is the frequency of the oscillation.

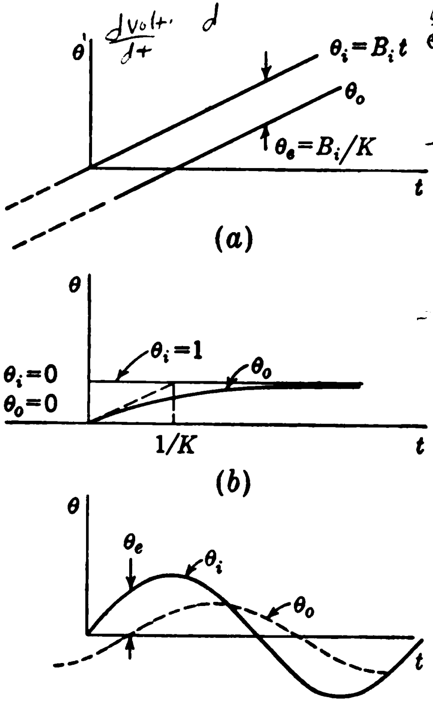


FIG. 1.2. Errors for standard inputs: (*a*) constant-velocity, (*b*) unit-step input, (*c*) sinusoidal input.

Ex.
θ_i = vel. at which volt[illegible]
[illegible] = vel. at which [illegible]
[illegible] =

For a system with inertia considered, an error is also required to cause the output to accelerate. With two specifications given, the response to a step input or a sine-wave input could be determined. For a more complicated system, the specifications for one form of input are still related to those for another form, although not directly by simple mathematics.

The form of the specification for a system depends upon the nature of the input encountered. An airplane camera mount, for example, must rotate the camera, with respect to the airplane, opposite to the rotation of the plane itself. A typical specification for the mount is the maximum allowable error when the plane is rolling sinusoidally through its typical amplitude at the natural frequency of the airplane in roll. A voltage regulator is often subject to having part of its load switched on or off the line. Such a sudden step change starts the natural transient of the regulator. A fast-acting linear regulator will overshoot in its correction, and the percentage overshoot may be specified. Similarly, the time required for the transient to die out to 5 per cent of the total change is often of interest. For a gun turret the angular rate of the sight in tracking the target is essentially constant, and therefore the error of the servo in driving the turret at this rate is specified. The rate of change of tracking rate may also be important; this would require that the error for accelerating the turret be specified.

---

[1] Refer to the terminology section at the front of the book for an explanation of the choice of symbols.

All these specifications are in the form of a ratio of an allowable error to a specified input. The use of a ratio implies that the ratio is constant, *i.e.*, that double error would result from doubling the specified input. However, this linear relationship does not extend over the whole range of operation. Actually, the servomechanism has maximum output limited by such things as the power-supply voltage for the amplifier and the saturation of the magnetic field of a motor. The linear analysis is carried out first, because it is approximately correct over the middle range of operation and it is very rapid, even for a complicated system, compared to a nonlinear analysis of a comparatively simple system. Upon completion of this analysis, however, the maximum values required of the components should be checked, in order to ensure that the results are valid.

The specifications of a servo can sometimes be relaxed by taking advantage of all information regarding the type of signal it must follow. A machine-tool control is an excellent example because the desired motion of the tool is known in advance. Knowing also the motor voltages required to drive the tool at various velocities permits a time program of motor voltage to be computed which would cause the tool to be positioned approximately. Measurement of the error between actual tool position and the desired position would then permit corrections to be made in the motor voltage. Such a control has been called "open cycle–closed cycle"[1] to emphasize that it operates partially without benefit of the measured error.

A servo also has specifications of size, weight, and complexity which limit the performance which can be achieved. The final design is, therefore, a compromise based upon many factors, rather than upon accuracy specifications alone.

[1] J. R. Moore, Combination Open-Cycle Closed-Cycle Systems, *Proc. IRE*, Vol. 39, No. 11, pp. 1421–1432, 1951.

1.4. Fundamentals of Linear Equations

A linear control system is one whose characteristics are independent of the amplitude of the signal in any part of the system. Great simplification in analysis is permitted for such a system because the effect of one input can be studied independent of all other inputs. Moreover all transient responses can be shown to be the sum of signals of the same mathematical form.

The control system which makes a gun turret follow the motion of a sight has other inputs, such as wind forces on the gun turret. Fortunately the effect of each input can be studied independently. The sight motion itself may often be approximated as a constant velocity of tracking the target plus an oscillation introduced by the gunner's failure to track perfectly. These two signals can be considered separately even though they are inputs from exactly the same point in the system. Such simplifications are possible whenever the characteristics of each component in the system are essentially independent of the total amplitude of the signal present. The most common limitation is that the signal be within the normal rating of the component; thus the output voltage of an amplifier must be within its limits or else an additional input voltage would not cause any additional output voltage.

When any part of the system is operating within its linear range, however, the relationship between its input signal and its output signal can be described by an equation in which all characteristics of the components, such as resistance and inductance of a generator field, are constant. The constant-velocity motion of the sight will correspond finally to constant current in a generator field, constant motor speed, etc., such that all equations are satisfied. The additional oscillatory motion of the sight will cause additional currents, velocities, etc., such that again all the equations are satisfied. Note that, for an input signal of a given form, the signals which result in all parts of the system are of the same form.

In addition to the steady action of input signals, however, another situation can exist. Consider the special case when all inputs are suddenly made zero at a time when the gun turret is moving, when capacitors are charged, etc. These conditions represent storage of energy which must be gradually dissipated before the system comes to rest. Similarly the presence of amplifiers and other energy sources in a system creates the possibility of signals building up rather than dying out. For almost any practical system, however, the amplifier settings must be such that the transients die out. The system is then said to be stable. Mathematically the time variation of any signal during such an interval can be shown to be expressible as the sum of terms of the form e^{st} and is called the transient response. Transients are created at almost any time a new input is impressed or removed, and again linearity of the system permits these to be studied independently.

1.5. Frequency Response

> *The term frequency response of a system refers to the change in magnitude and the shift in angle of sinusoidal signals of various frequencies in being transmitted through the system. The characteristics may be calculated or measured and provide a picture of the system which is useful in avoiding instability.*

The frequency-response method (see Chap. 5) avoids the problem of finding transients by determining only the steady response of the system to sinusoidal inputs. For convenience the components are considered individually when possible; each one is described by the change in magnitude and the shift in angle of a sine wave in being transmitted from input to output of that component. This information for sine waves of various frequencies is called the "frequency response" of that component. If two components act in sequence in a control system, the output of one is the input to the other. The dividing line between components is made in such a way that their characteristics are not changed by making the connections; hence the amplitude of a sine wave in going through a series of components is changed by the product of the factors for amplitude change in each component, and the phase angle is shifted by the sum of the phase-angle changes for each component. Starting with an assumed error signal, the effect of each component in sequence can be added until the output signal is determined. The error signal for a simple control system is the difference between the input and the output; or the input is the sum of the error and the output. With error and output both known, the input which would be needed for the originally assumed error is simply the sum of the sine waves for the error and the output. The transmission of sine waves through components or the combination of sine waves at junction points in the block diagram can thus be easily calculated. The frequency-response description is sufficient for many purposes, such as setting the limits for stable operation.

1.6. The Root-locus Method

> *A transient response can be expressed as the sum of terms of the form e^{st}. The values of s are roots of the characteristic equation of the system. The root-locus method determines these roots graphically using simplifications suggested by the form of the equation for a control system.*

A control system is somewhat unique in that the sequence of cause and effect is very clear-cut. Thus the error angle between the sight and the gun turret is measured electrically; the error voltage is amplified and impressed across the motor, which drives the turret in the direction to reduce the error. The frequency-response method takes advantage of this sequence by describing each component by its transmission of a sine wave. In the root-locus method the component is described by the roots of the characteristic equation of that component. Such an equation is simple, and the roots are comparatively easy to find. When the components are connected together, but with the amplification of the error signal kept low, the individual transient of each component is essentially unaffected by the transients of other components. As the amplification is increased, however, the interaction is increased. Mathematically the roots of the characteristic equation of the entire system shift from their original values set by the individual components. A root-locus plot determines this shift in all roots as a function of amplification. Trial and error is required in making the plot, but approximate results can be determined very rapidly. Later analysis will show that the effect of any other parameter of the system, such as a capacitance in a lead network or the spring constant of a shaft, can be made the variable for a root-locus plot. The method involves only one basically

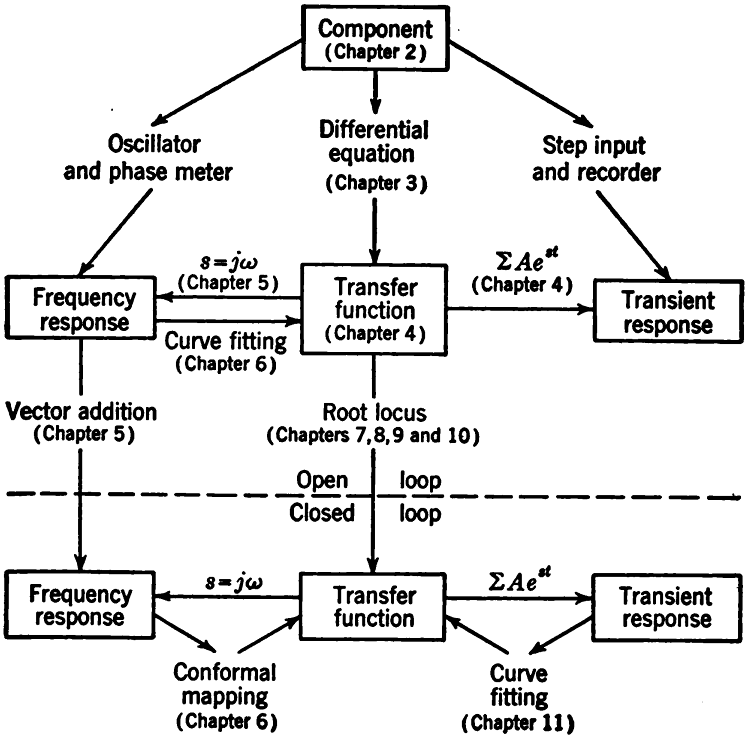


FIG. 1.3. Interrelation of methods.

new idea, which permits greater use of much valuable theory already developed: it will be shown that s in general is a complex number which can be plotted in a plane. This plane, which is the one used for a root-locus plot, is also the one which permits specifying the characteristics possible for any linear system in the simplest possible way, as a set of points. The frequency response of the system could be computed at any stage of the analysis at which it is of interest. The interrelation of methods is shown in Fig. 1.3.

1.7. Nonlinearities

Test results, as well as two separate methods of analysis of nonlinear systems, indicate that many of the concepts of linear systems can be extended to cover common nonlinearities. One method is easily applied to a system of any complexity as long as only one nonlinearity exists; the other can handle several nonlinearities as long as only two energy storages are involved.

Except in certain special cases the exact analysis of a nonlinear system is very difficult. Fortunately, however, approximate results can be obtained fairly readily in many common problems. If the only nonlinearity is backlash in a gear train, for example, with the load torque small compared to the motor torque, the effect can be described in terms of transmission of sine waves of various amplitudes. As the amplitude of oscillation of the input shaft is decreased, the ratio of output motion to input motion is decreased and the effective angle of lag of the output shaft behind the input shaft is increased. This characteristic used in conjunction with the frequency response of the linear portion of the system often determines the frequency and amplitude at which a steady oscillation will be maintained. Tests verify this oscillation, in which the output shaft motion is usually comparable to backlash angle.

If only two energy-storage elements are involved in a follow-up system, then the angle and velocity of the output shaft completely describe the condition of the system at any instant. A plot of angle versus velocity for any given nonlinearity can be made as a family of curves for different amplitudes. The time required for a motion from one operating point to another is usually not known, but effects such as coulomb friction are easily introduced because velocity is known. If no input motion is involved, then relay action can be introduced, because a line can be drawn on the plot to show where contact is made.

For a given system nothing, of course, can replace a direct test. A common procedure where necessary is to simulate the portions of the system not available and connect them into a complete system having as many actual components as possible. The methods of analysis provide concepts which are useful in understanding test results.

CHAPTER 2

DESCRIPTION OF TYPICAL COMPONENTS

This chapter describes the important characteristics of a few typical components used in control systems. The measurement device of a position-control system, for example, must be accurate, because the accuracy of the system cannot be better than that of the measurements. The amplifier must be flexible, because the measurement device and the actuator are usually chosen first, with the amplifier used to accept the signal from one and supply the power to the other. The actuator often must be fast, because the output cannot be driven any faster than the actuator is capable of driving it. The importance of good components is much greater than the treatment given in this book would indicate. This unfortunately must be so because the design of a good component generally involves an intensive study of a subject and the subjects are wide and devious.

The description of a component in this chapter is carried as far as its differential equations of motion. Other descriptions involve some form of solution of that equation, and these topics are treated later. In some cases, such as the torque-speed characteristics of an induction motor, it is more convenient to present the information in the form of curves from which linear approximations are made.

2.1. Angle-measuring Devices

> *Synchros are electromagnetic devices in which the transformer coupling between rotor and stator coils is a function of mechanical position of the rotor. Synchros are used to indicate the position of one shaft with respect to a remote shaft or to drive a shaft which is lightly loaded.*

The simplest type of device for giving a voltage proportional to a displacement is a potentiometer like that shown in Fig. 2.1. The motion of the slider across the wire-wound potentiometer causes it to pick up a voltage which is essentially proportional to the distance moved. Standard linear potentiometers are usually rated as having a maximum deviation from linearity, compared to the total resistance, of 10 per cent.

Precision potentiometers, however, are rated as being linear to within 0.1 per cent or better.

The relative position of two sliders x_1 and x_2 is obtained by comparing the voltages between the sliders. Note that, as long as the same voltage is impressed across both potentiometers, this voltage may change without introducing an error in the measurement. Each slider voltage may be expressed in terms of a fraction of the total resistance. Most potentiometers are wound in a circle arc of about 270°, although some are made in a helix of 10 revolutions.

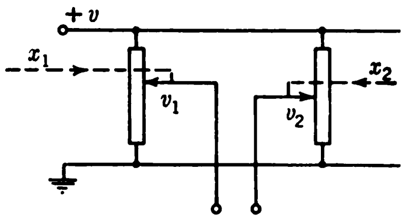


FIG. 2.1. Potentiometer bridge circuit.

For complete freedom of angular rotation synchros are usually used. A synchro is one form of an electromagnetic device excited with alternating current, as shown in Fig. 2.2. The rotor is excited with alternating current, which sets up a pulsating magnetic field along the axis as shown. The stator has three windings 120° apart, as indicated by the axes of the coils A, B, and C. The magnitude and polarity of the voltage induced in each coil depends upon its position with respect to the rotor field. In particular, the voltage induced in each winding can be found by projecting a vector for the field onto the axis of each of the coils. This construction gives a voltage which is proportional to the cosine of the angle between the field and a coil axis and assumes that the pattern of field intensity versus angle around the rotor is sinusoidal in shape.

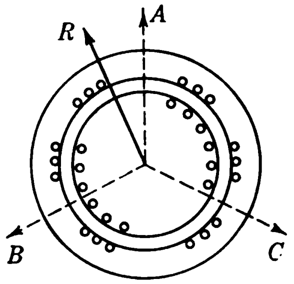


FIG. 2.2. Schematic diagram of a synchro.

The three lines from the stator windings may be connected to windings of a similar synchro. The voltage impressed on each coil of the second synchro is thus the same as that induced in the corresponding coil of the first synchro. The currents in these coils must set up a pulsating magnetic field in the second synchro; this field is in the same position with respect to the coils as in the first synchro. The synchro whose rotor was excited is called the "generator," and the one in which the stator coils set up the field is called the "control transformer." If the rotor of the control transformer is turned 90° with respect to the magnetic field, no voltage will be induced in the rotor coil. But for deviations of the rotor from this position, voltages will be induced which are proportional to the angle of rotation, with polarity corresponding to the direction of rotation. This combination of two synchros is thus an error-indicating device between the rotor axis of the generator and the cross axis of the control transformer. In a typical application this error voltage of about 1 volt/deg is used to control the output shaft of an actuator to which the control transformer is attached; the generator is attached to the input shaft whose motion is to be followed.

A follow-up system complete in itself and capable of driving light loads can be made by exciting the rotor of the second synchro, which is then called a "synchro motor." A torque acts on the rotor tending to make the axis of the rotor coil line up with the field set up by the currents in the stator coils. For a synchro motor 2 in. in diameter, the torque developed is about $\frac{1}{2}$ in.-oz/deg from neutral.

Basically a synchro unit can be used in any of the three applications: generator, control transformer, or motor. Design features in the three units do differ, however, for the following reasons: The generator has a larger rotor winding to keep down the i^2R loss in this field, which indirectly magnetizes both units. The control transformer has many turns in the rotor winding in order to have a large voltage output. The motor has some sort of friction damper on the rotor in order to prevent oscillation.

A differential synchro has three windings on both the rotor and the stator. When these windings are in line, the voltages induced in the rotor windings are the same as the voltages impressed on the stator windings. But if the rotor is mechanically rotated through some angle such as 30°, the rotor voltages are changed in exactly the same way as if the rotor of the generator had been rotated through 30°. The differential synchro, as its name implies, is therefore a means of adding an angle, as must be done in a remote-controlled gun turret to correct for the difference in positions of sight and gun. The generator in the sight excites the differential synchro in the computer. The computer rotates the differential synchro through the lead angle so that the signal to the gun-turret servo is the direction in which the turret should be pointed.

Greater accuracy can be achieved by use of a "two-speed" synchro system. The input shaft drives a synchro directly and also a synchro whose rotation is geared up by a ratio such as 31:1. The stator windings are connected to two corresponding synchro control transformers. These signals are both used as inputs to an amplifier whose input circuit follows the direct-connected synchro for large errors, but follows the 31-speed synchro when the error is small, so that the final accuracy is 31 times as great as it would have been with a single synchro system.

2.2. Small-motion Pick-offs

Many applications require the measurement of a very small motion of the order of 1/100 *in. or even* 1/100,000 *in. The applications are common because many devices for measuring pressure or temperature use a mechanism which gives a small displacement propor-*

tional to the quantity being measured; then this displacement is measured electrically in order to give the output signal.

The pressure difference between the inside and outside of the bellows in a pressure-measuring device causes motion against the spring action of the bellows so that the displacement is proportional to the pressure difference. A temperature meter might be based on the measurement of the displacement of a bimetallic strip. The acceleration of an object can be measured by the displacement of an inner mass with respect to an outer case, which is proportional to the spring force needed to accelerate the inner mass. These devices indicate the importance of measuring small motions.

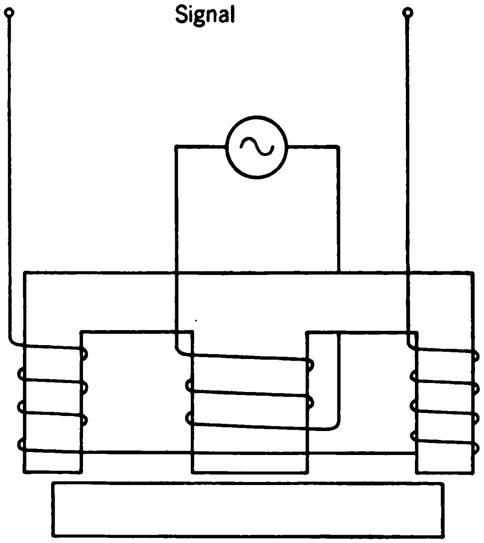


FIG. 2.3. Variable-reluctance pick-off.

One of the most common devices to measure small motions is the "E pick-off," so called because of the shape of the lamination shown in Fig. 2.3. The coil on the center leg is excited from an a-c source. The resultant a-c field crosses the gap through the movable armature piece and returns through the two outside legs. If the armature is centered, the flux splits evenly; the voltages induced in the two outside coils are equal and opposite. With the armature shifted to the left, the flux returning through the left leg is greater than that through the right, making the net voltage have the polarity of the left coil. Similarly, if the armature is shifted to the right, then the net voltage has the polarity of the right coil. Note that no windings are needed on the moving element but that magnetic forces may act on the element.

An inductive type of pick-off is shown in Fig. 2.4. The two C-shaped iron cores are set with pole faces in line and a small gap between them. The four coils are excited with a-c voltages and set up an a-c field, which circulates around the two C cores as shown by the dotted lines. A coil of fine wire in the shape of a thin disk is mounted on an arm and inserted in the gap. When the arm is centered, the lines of flux have no net linkage with the coil, and no voltage is induced. But as the coil is displaced to one side or the other, a voltage is induced whose polarity depends upon the direction of motion. The inductive type of coil has an infinitesimal coercion force because the force depends upon the current in the coil, which may be very small, as when the load is the grid of a vacuum tube.[1]

Capacitance circuits in general have the highest possible sensitivity in terms of volts per inch because the applied voltage can be high although the impedance of the capacitors is figured in megohms. A bridge circuit uses the capacitance of a movable plate between two fixed plates. The output voltage of the bridge is zero when the arm is centered; moving the center plate up increases the capacitance of the top section and decreases that of the lower section. Greater sensitivity can be achieved by having the capacitance change unbalance a circuit which is resonant at the excitation frequency.

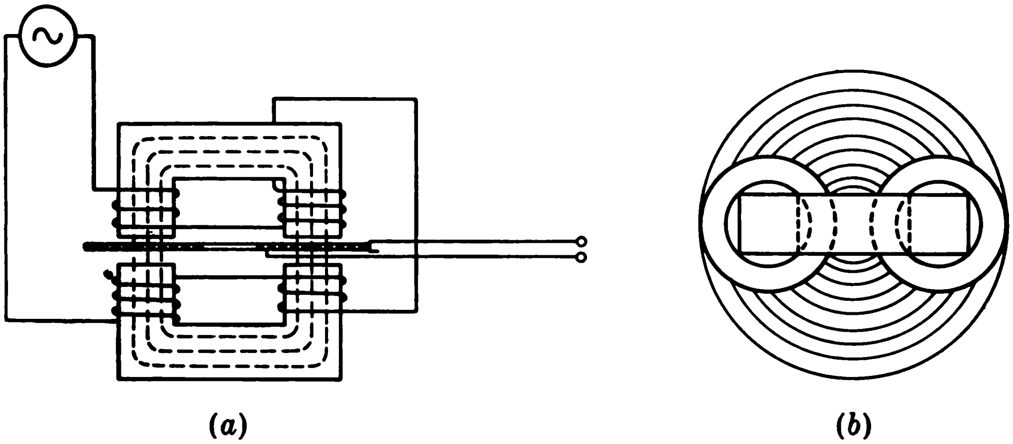


FIG. 2.4. Inductive type of pick-off: (*a*) side view; (*b*) top view.

2.3. Amplifiers

> *A-c amplifiers have the advantage of ease of coupling and stability of gain. D-c amplifiers are subject to drift but are desirable because stabilization networks for d-c signals are simpler.*

An input signal may be alternating current from a synchro or direct current from a potentiometer; the required output may be direct current for a solenoid or alternating current for an induction motor. The amplifier is almost certain to have some circuits included for the purpose of stabilizing the system. These stabilizing networks themselves may be for either a-c or d-c signals, except that the d-c types are far simpler and are used when possible. With this broad scope of material to consider, this section is concerned simply with typical amplifier characteristics. Conversion devices from direct to alternating current (modulators) or from alternating to direct current (demodulators) are considered in the next section.

---

[1] This design was developed by Ward A. Harman, now at Stanford University, while at the Aerophysics Laboratory, North American Aviation, Inc., in 1948.

A typical a-c amplifier to drive a two-phase induction motor from the error signal of a synchro involves some voltage-amplification stages followed by a power stage (Fig. 2.5). The voltage at the plate of the first tube can be coupled to the grid of the next tube through a capacitor which charges itself to the average level of the first plate so that the d-c level of the second grid is near ground. The potentiometer is a simple means of over-all gain control. The plate load of the second triode is a transformer having a center-tapped secondary because the output stage is push-pull. Thus the grid of the one tube is positive during one half cycle, while the grid of the second tube is negative. The voltages at the ends of the primary are thus lowered alternately. The secondary is connected to one phase of the induction motor while the other phase has a fixed a-c voltage impressed across it. Often the impedance of the motor is far less than the internal impedance of the amplifier; the transformer ratio permits transmission of maximum power to the load.

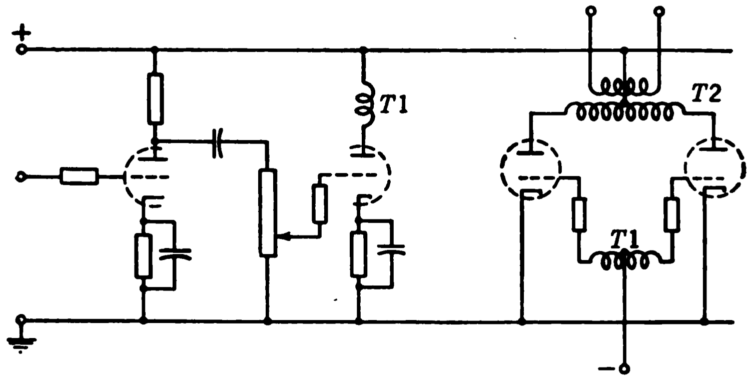


FIG. 2.5. Typical a-c amplifier.

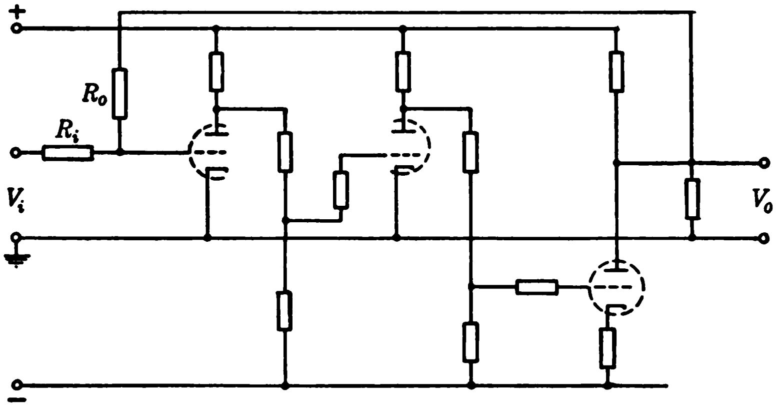


FIG. 2.6. Typical d-c amplifier.

A typical d-c amplifier is shown in Fig. 2.6. The voltage across the output is made proportional to the input voltage by having the load voltage fed back to the input. The error between the input and output voltage is greatly amplified to drive the output. Note that this feedback amplifier operates on the same principle as a control system, in that the controlling signal is the error between the input and the output. The amplifier has an odd number of stages so that the output is negative when the input is positive. The grid voltage is the average of these two voltages if the coupling resistors R_i and R_o are equal. One problem arising in d-c amplifiers is that the output of one stage is at plate potential and must be shifted down to the level of the grid of the next stage. The method shown uses a voltage divider whose lower point is at a negative potential. This necessarily involves an attenuation of the signal by the voltage-divider ratio.

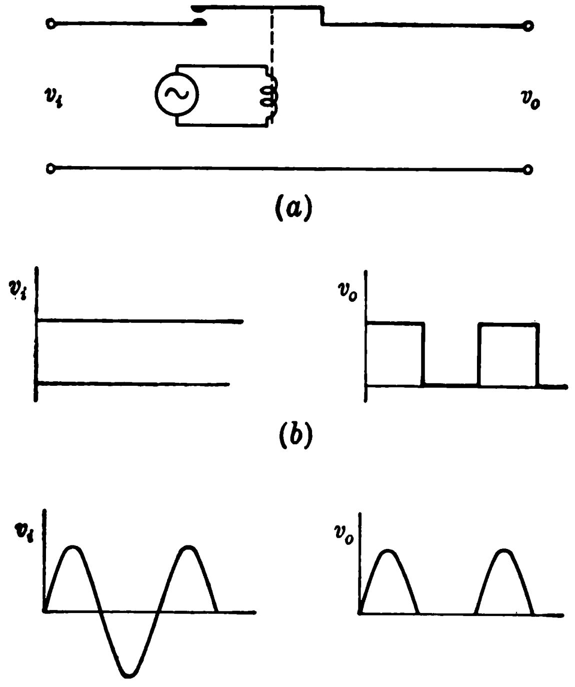
A major problem in d-c amplifiers is drift of the output voltage independent of the input voltage signal. A major cause of this drift is "cathode potential," which is due to the finite velocity with which electrons are emitted from the cathode. This effect is equivalent to an impressed grid voltage, and it changes with the temperature of the cathode. If the drift effect is too great, the amplifier should convert the d-c signals to an a-c signal, amplify, and then convert back to a d-c signal, using devices discussed in the next article.

2.4. Modulators and Demodulators

Conversion between a-c and d-c signals can be achieved by a switch which is alternately opened and closed in synchronism with a given frequency. The switch serves as a "modulator" in producing a square-wave output by alternately short-circuiting a d-c signal; it serves as a "demodulator" by transmitting only half waves of an a-c signal of the given frequency.

The synchronous switch is shown in Fig. 2.7*a*. The electromagnet is excited from the given a-c supply so that the contacts are closed for alternate half cycles. If the device is used as a modulator (Fig. 2.7*b*), the input d-c signal is transmitted during half cycles, producing a square-wave output at the frequency of the modulator. The same switch can serve as a demodulator (Fig. 2.7*c*) of a signal having the given frequency. The circuit is closed only during the positive half cycles of the input signal, thereby producing a d-c output signal. A capacitor across the output would provide filtering action with its charging rate determined by the resistance of the source and its discharge determined by the shunt resistor.

A synchronous switch can be obtained by using rectifiers, as shown in Fig. 2.8*a*. During the half cycle in which the point C is driven positive



(a)

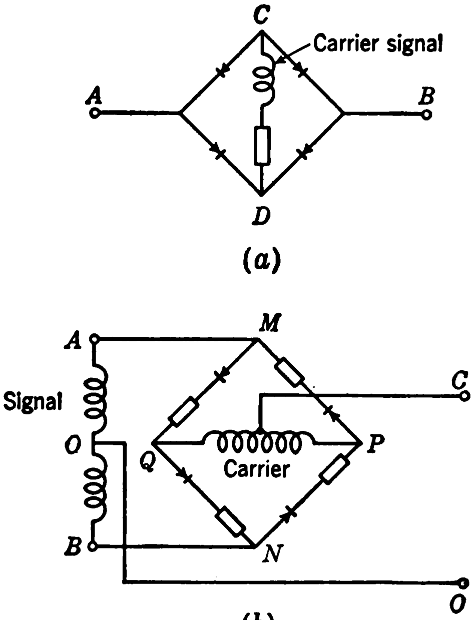
(b)

(c)

Fig. 2.7. (a) Mechanical synchronous switch; (b) modulation; (c) demodulation.

with respect to point D by the carrier winding, current through the rectifiers establishes a path from A to B as long as the net current through each rectifier is forward.

Full-wave demodulation can be achieved with a ring demodulator, as shown in Fig. 2.8b. During the half cycle in which point P is positive with respect to Q, current flows from P through M to Q. The point A is then effectively connected to the output C. During the next half cycle, point B is effectively connected to point C.



(a)

(b)

Fig. 2.8. (a) Synchronous switch; (b) ring demodulator.

Conversion between a-c and d-c signals is useful in control-system amplifiers for a variety of reasons. An error signal is more easily modified by transmission through a network if the signal is a d-c voltage proportional to the error. If the error signal is originally alternating and the final output required is alternating, the use of d-c types of networks then requires a demodulator and modulator. Fortunately these devices are available as plug-in units.

2.5. D-C Motors

> *A d-c motor has stall torque and maximum speed proportional to the impressed voltage. The decrease of torque with speed is linear because it corresponds to the decrease in iR drop in the armature with increasing counter emf.*

Of the various types of d-c motors, the type with the fixed shunt field is the simplest to analyze. Its most common application is in control systems where the power required is sufficient to justify the use of a generator to control the voltage applied to the armature circuit. The motor can be represented schematically as shown in Fig. 2.9.

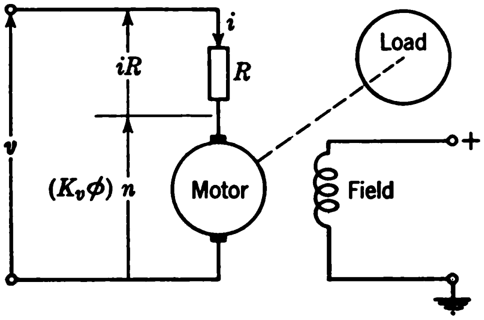


FIG. 2.9. D-c shunt-motor.

The applied voltage is equal to the sum of the iR drop in the armature and the counter emf generated in the rotor. This generated voltage is proportional to the flux ϕ, which is constant for a shunt motor, and to the motor speed n.

$$v = iR + K_v\phi n \tag{2.1}$$

where K_v is a constant. The torque m developed by the motor is proportional to the armature current i and to the flux.

$$m = K_m\phi i \tag{2.2}$$

where K_m is a constant.

The division of power input between output and losses can be found by multiplying the voltages in (2.1) by current:

$$vi = i^2R + K_v\phi ni \tag{2.3}$$

If mks units are used, electrical power and mechanical power need no conversion factor, and the constants K_m and K_v are identical; in any system they are proportional, as can be seen from the alternate ways of expressing power output.

$$K_v\phi ni = K_m\phi in = mn \tag{2.4}$$

Since the iR drop falls off linearly as the speed increases and since torque is proportional to current, the torque-speed curves are a series of straight lines, as shown in Fig. 2.10.

The stall torque for full voltage v applied is $K_m\phi(v/R)$, and the top speed with no torque required is $v/K_v\phi$. As the voltage is reduced, both stall torque and top speed fall off in direct proportion. The loss of developed torque with the increase in speed is referred to as the damping characteristic of the motor. In the study of the dynamics of the complete system to be considered later, this torque decrease is the same as if a dashpot were present which absorbed a torque proportional to speed.

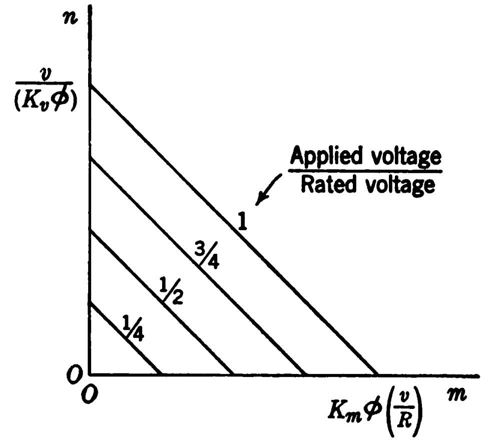


FIG. 2.10. Torque-speed characteristics of d-c shunt motor.

The differential equation of the motor relating applied voltage to resultant speed can be written assuming some particular form of load such as an inertia J. The torque developed by the motor then produces an angular acceleration:

$$m = K_m\phi i = J\frac{dn}{dt} \tag{2.5}$$

Substituting in (2.1) for current in terms of acceleration from (2.5) yields

$$v = J\frac{dn}{dt}\frac{R}{K_m\phi} + K_v\phi n \tag{2.6}$$

Combining the constants shows that (2.6) is of the form

$$v = K_m\left(1 + T_m\frac{d}{dt}\right)n \tag{2.7}$$

The solution of this equation and the significance of the constant T_m are considered in Art. 3.2.

D-c motors are also used in small sizes in the form of split-series motors and field-controlled motors. In each case the flux intensity varies, making the characteristics of the motor nonlinear. They are rather readily determined from the equations already given, but usually the friction torques and effects of saturation are so pronounced that the measured motor characteristics are not what would be expected with these effects neglected.

2.6. A-C Motors

Induction motors with one phase fixed have stall torque proportional to the voltage applied to the variable phase, but the top speed is essentially constant. These characteristics can be obtained by use of symmetrical components.

The two-phase induction motor is very commonly used in applications in the range from 1 to 50 watts output. A typical circuit arrangement is shown in Fig. 2.11. One phase of the motor is supplied from a fixed source with the capacitor in series to obtain a phase shift of approximately 90°. The other phase is driven by a push-pull amplifier so that the polarity of this field can be reversed. The torque-speed curve for such a motor for full voltage across the variable phase is compared with that of a standard induction motor in Fig. 2.12.

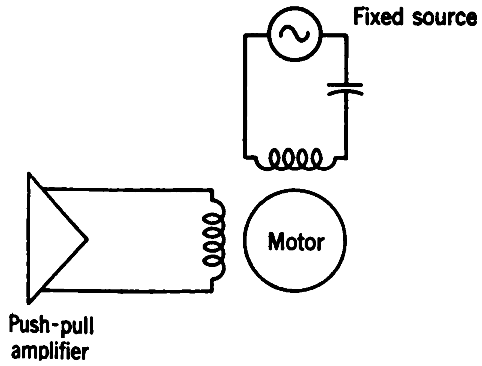


FIG. 2.11. Two-phase induction motor.

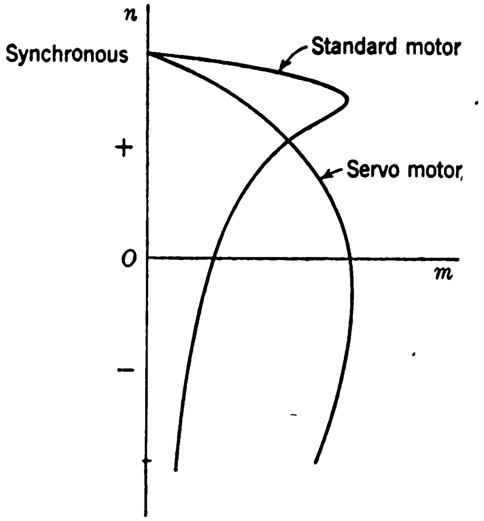


FIG. 2.12. Comparison of induction-motor torque-speed characteristics for servo motor and standard motor.

The torque-speed curves actually have the same shape; the difference is just that the peak torque of the servo motor is made to occur near the stall point rather than at about 20 per cent slip from the standard motor. The reason, of course, is that rapid acceleration is the prime feature of the servo motor, rather than efficiency

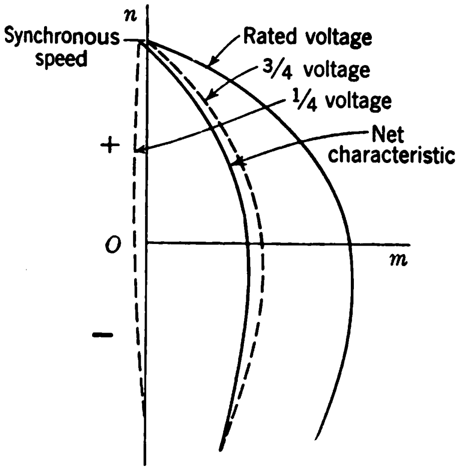


FIG. 2.13. Torque-speed characteristics for half rated voltage on control phase.

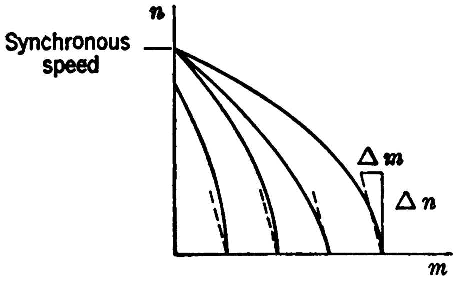


FIG. 2.14. Linear approximation of induction-motor characteristics.

as for the standard motor. The curves are extended into the region where torque is opposite to existing speed.

The characteristic of the motor with only partial voltage across the variable phase is obtained by considering symmetrical components. If the variable phase has one-half rated voltage, the resultant field pattern in the motor is the same as if there were a field of three-fourths rated value rotating in the positive direction and a field of one-fourth rated value rotating in the negative direction. The torque for each field varies as the square of its strength. The net torque from these two oppositely rotating fields can be obtained by superposition (Fig. 2.13).

The complete set of curves for variable voltage on one phase is shown in Fig. 2.14; a linear approximation to these curves for low-speed operation is shown by dotted lines. The motor is treated the same as a d-c motor except that the damping, $\Delta m/\Delta n$, is nearly zero.

2.7. Solenoids and Clutches

> *A solenoid is an electromagnetic device, usually spring-loaded, to give a displacement proportional to current. A magnetic clutch transmits torque proportional to the magnetic field set up in a suspension of iron filings between the clutch faces.*

A solenoid is an electromagnetic device which can produce a large force but is limited to a small range of motion. In either case, however, the solenoid is almost sure to have some spring-loading effect, as shown in Fig. 2.15.

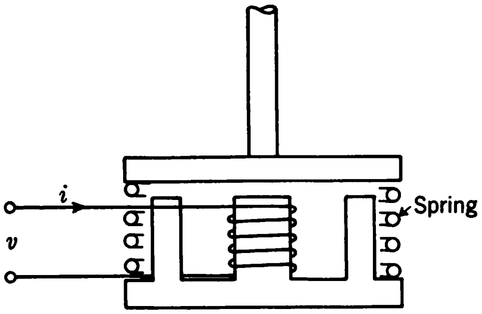


FIG. 2.15. Spring-loaded solenoid.

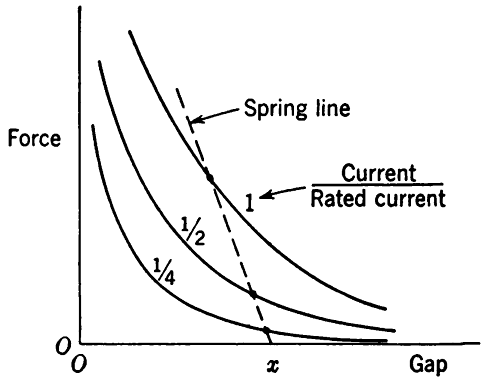


FIG. 2.16. Force-displacement curves for solenoid.

The relation between the input voltage and the output displacement is quite complicated, particularly because of the nonlinear basic force curves of a solenoid, shown in Fig. 2.16. For constant current in the coil, the force varies inversely as the square of the gap up to the point of saturation of the magnetic circuit. For a given gap the force varies as the square of the current. The armature has a strong tendency to snap in to the position of zero gap; therefore, the spring must have a neutral position and stiffness sufficient to prevent this action. Thus in Fig. 2.16 at each operating point the spring force increases faster than the magnetic force for a decrease in gap, producing a stable system.

The solenoid may be operated with a neutral position of the load device set at the position of the solenoid at the point for half current. Increase of current then produces a positive motion and decrease in current produces a negative motion of the load.

The differential equations of the solenoid relating displacement to voltage are nonlinear for large displacements but can be expressed in linear form for small deviations from the neutral position. The equation of voltage equilibrium is

$$v = iR + L_0 \frac{di}{dt} + B_0 l \frac{dx}{dt} \tag{2.8}$$

In this equation, L_0 is the inductance of the coil at the neutral position, B_0 is the flux density at the neutral position, and l is the length of the coil. The net force acting is proportional to change in current from the bias value; the load will be considered to be only the spring force and inertia.

$$f = Ci = Kx + J \frac{d^2x}{dt^2} \tag{2.9}$$

Substituting for the current from (2.9) into (2.8) gives

$$v = \left(R + L_0 \frac{d}{dt}\right)\left[\left(\frac{1}{C}\right)\left(K + J \frac{d^2}{dt^2}\right)x\right] + B_0 l \frac{dx}{dt} \tag{2.10}$$

The magnetic clutch is a device, developed at the National Bureau of Standards, in which the clutch faces of the input and output shafts do not make mechanical contact at any time. Instead, the oil between the faces has fine iron particles suspended in it so that the mixture congeals when a magnetic field is applied. This congealing is so strong that torque can be transmitted between faces with a shear force of 35 psi. Moreover, the torque transmitted is approximately a linear function of the magnetic field strength developed, a property which makes the device very desirable for use in control systems. The speed of response is very great because the acceleration is limited only by the inertia of the load; the power source itself is running at constant speed.

The main problem in using the clutch is the dissipation of power when the clutch is slipping, because the power involved is the torque being transmitted times the slip speed. Two-direction control may be achieved by use of two clutches operating off shafts rotating in opposite directions.

Each clutch always transmits some torque, so that the backlash in the gear train is eliminated. This makes torque variation a smooth function of current variation in the clutch coils and solves a serious problem in precision servos.

2.8. Hydraulic and Pneumatic Components

Hydraulic systems have many intercoupling effects because of the many interrelations between pressure and flow. A complete control system can be made using pneumatic components alone.

Hydraulic components frequently must be studied in combination rather than as individual components. Thus consider the case of a hydraulic cylinder controlled by a pilot valve, as shown in Fig. 2.17. The current in the solenoid produces a corresponding opening of the ports of the pilot valve. The figure shows the ports half open with the left side of the cylinder having high pressure and the right side ported to the low-pressure reservoir. The displacement of the piston is proportional to the quantity of oil which has moved around the circuit. The velocity of the piston requires a velocity of oil flow through the half-open port which causes a pressure drop at this point. The result is that the pressure in the cylinder falls off with velocity.

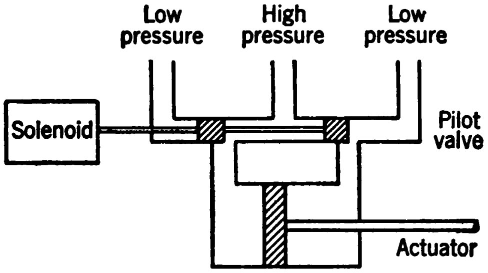


FIG. 2.17. Hydraulic valve and actuator combination.

Note that the peak pressure is always that of the high-pressure source but that the top velocity is roughly proportional to the valve opening. This is not a linear system like the d-c shunt motor, which has parallel torque-speed lines, in that top pressure does not decrease with port opening.

For the ideal case, the oil would be circulated around the loop with no leakage so that the motion of the actuator would be exactly proportional to the quantity of oil flow. Actually a pressure must develop to produce the force required by the load for a given speed; this pressure causes a leakage of oil past the actuator. Pressure causes expansion of the tubing, permitting a storage of oil. For long lengths of tubing the distributed effects of pressure drop along the tube and expansion of the tube require that the tubing be considered by equations similar to those for electric transmission lines. One method of studying hydraulic systems is by an electric analogy.

Pneumatic devices are in many operating characteristics quite similar to hydraulic devices, but the applications are more often on low-power systems. A significant difference is that the exhaust air can be ported into the atmosphere; no reservoir is needed. This is very convenient on instrument-control systems such as gyroscopic devices. The compressibility of the air makes a difference in the response characteristic of a pneumatic device compared to a hydraulic device, where the only similar effect was due to expansion of the tubing under pressure.

Pneumatic devices include pick-offs, amplifiers, and actuators, so that they may be used to form a complete control system, as illustrated in Fig. 2.18. The pick-off is the vane, which can be moved across the two jets. The jet which is more nearly blocked allows less flow through its restrictor and causes a higher pressure to be developed on one side of the small piston. This piston drives the other valve, which ports the high-pressure air into the actuator, which drives the jet block back to a neutral position with respect to the vane. The entire system is thus pneumatic.

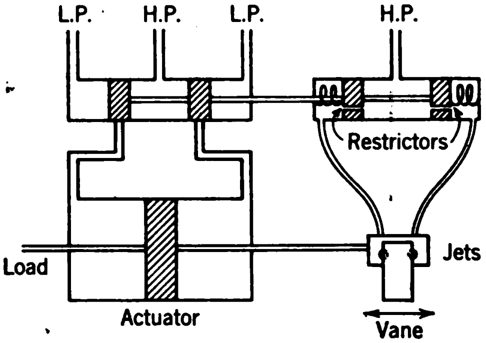


FIG. 2.18. Complete pneumatic system.

2.9. Choice of Gear Ratios

Maximum acceleration of the load is achieved by making the inertia of the load as seen at the motor equal to that of the motor. Static friction and gear inefficiency require that the gear ratio be increased.

A motor design for large power in a small size leads to high top speeds and low torque. Many loads, however, such as a radar dish, require high torques at comparatively low speeds. The problem then is to select the gear ratio for maximum acceleration of the load. The quantities involved in the problem are shown in Fig. 2.19.

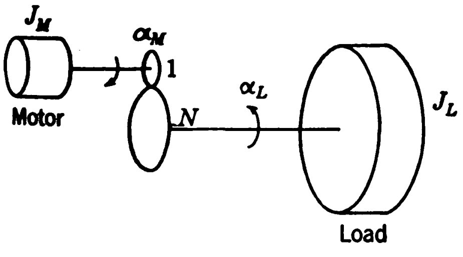


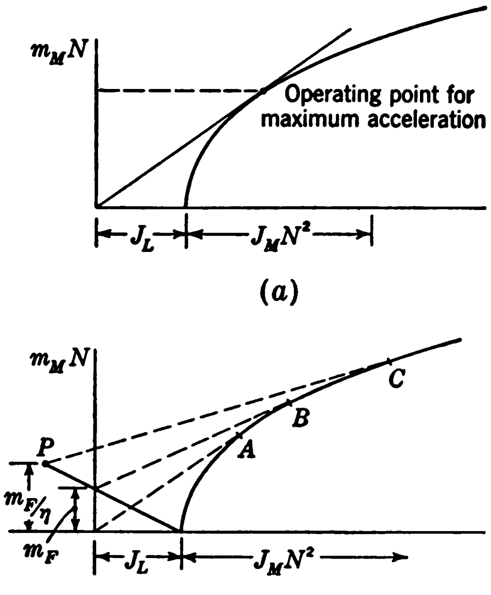
FIG. 2.19. Problem involving choice of gear ratio.

The terms J_M and J_L refer to the polar moments of inertia of the motor and the load, respectively. The gear ratio is N, and η is the efficiency of the gear train, which is defined as the ratio of torque actually delivered to that for an ideal gear train. m_M is the motor torque, and m_F is the static friction at the load. The torque equation expressed as the balance of torques at the load and is then

$$(m_M - J_M\alpha_M)N\eta = m_F + J_L\alpha_L \tag{2.11}$$

where α_M and α_L are the angular accelerations of motor and load. Divide the equation by the efficiency η and transfer all inertia terms to the right, using $\alpha_M = N\alpha_L$.

$$m_M N - \frac{m_F}{\eta} = \left(J_M N^2 + \frac{J_L}{\eta}\right)\alpha_L \tag{2.12}$$



(a)

(b)

FIG. 2.20. (a) Torque-inertia curve versus gear ratio; (b) effect of friction and gear inefficiency.

For convenience, consider first the special case when the static friction is zero and the efficiency $\eta = 1$. The plot of torque and inertia as a function of gear ratio is then as shown in Fig. 2.20*a*. The acceleration is the ratio of torque to inertia, which is the slope of a line from the origin to the curve at any point. This acceleration is a maximum at the point where the line is tangent to the curve. This result is equivalent to having the ratio of inertia to torque at a minimum. The latter statement is more easily used in setting the derivative with respect to N equal to zero in order to find the optimum value of N.

$$\frac{d}{dN}\left(\frac{J_L + J_M N^2}{m_M N}\right) = \frac{d}{dN}\left(\frac{J_L}{N} + J_M N\right)\frac{1}{m_M} = \left(\frac{-J_L}{N^2} + J_M\right)\frac{1}{m_M} = 0 \tag{2.13}$$

Note that the condition for maximum acceleration in the ideal case is that the inertia of the motor referred to the load should equal the inertia of the load.

Consider now the effect of static friction and gear efficiency, as shown in Fig. 2.20*b*. Static friction is subtracted from the motor torque, so that the acceleration with this effect included is the slope of a line from this offset point on the vertical axis to any point on the curve. The effect is to require a higher gear ratio than otherwise. The effect of gear efficiency, as seen in Eq. (2.12), is to increase the effect of static friction and the load inertia. The acceleration should now be figured as the slope of a line from the offset point P to any point on the curve.

The acceleration falls off rapidly as the gear ratio selected is less than the optimum but remains reasonably constant for gear ratios higher than the optimum. Thus allowance can be made for uncertainty in the value of constants by setting the gear ratio high. The ratio cannot be so high, however, that the load cannot reach the top speed required of it. If this limit is reached, acceleration must be sacrificed or a higher power motor selected.

2.10. Summary

> *The differential equation relating the input and output of a component completely describes that component for the purpose of a linear analysis of dynamic behavior.*

The typical components considered in this chapter demonstrate the wide range of knowledge necessary for design of all of them. Fortunately, the characteristics which are important for the analysis of dynamic behavior are limited. The torque-speed curves for a motor, for example, define the behavior of that motor. For a linear analysis, the approximations to the curves for a d-c shunt motor and an a-c induction motor give the same form of differential equation and thus the subsequent analysis is the same. The analysis of a hydraulic or pneumatic system is the same as that for an electrical system after the linear approximations are made.

The complete design of a control system has many more problems than the dynamic analysis of a linearized system. The determination of the requirements of the control system, the measurement or analysis of the given components, and the testing of the complete system are other aspects of the problem; skill in handling these aspects is more thoroughly acquired by experience than by theory. The foundation upon which to build this experience, however, should be a thorough understanding of the fundamentals.

CHAPTER 3

FUNDAMENTALS OF TRANSIENT RESPONSE

The purpose of this chapter is to study the physical concepts and mathematics needed for the analysis of transients in linear systems. For clarity these concepts are developed in terms of mechanical systems analogous to electrical control systems. Three such analogous mechanical systems are used: (1) spring and damper; (2) spring and mass; (3) spring, damper, and mass. The response of a system to an initial displacement is determined first in each case by physical reasoning. The mathematical description of the response is then shown to be, in the respective cases, (1) $e^{\sigma t}$, (2) $\cos \omega t$, and (3) $e^{\sigma t} \cos (\omega t + \phi)$. The simplest procedure for solving is to assume a solution of the form e^{st} in each case, because an algebraic equation is then obtained in which s replaces the operation d/dt of the differential equation. This general form of solution is applicable to all cases, although the value of s changes from a negative real number in system (1) to $j\omega$ in system (2), and finally to a complex number $\sigma + j\omega$ in system (3).

3.1. Mechanical Analogy for a Control System

A mechanical spring-mass system is analogous to a simple control system. Damping torques on the output shaft produce the same effect as the motor characteristic of developing less torque with increasing speed.

A mechanical analogy for a simple control system is shown in Fig. 3.1. In the control system, the error angle θ_e between the input and output shaft angles is measured and amplified. A proportional voltage v is impressed across a motor, which in turn produces a proportional stall torque m. In the mechanical system all of this is analogous to the spring itself. The error angle corresponds to the angle of twist of the spring, and the spring produces a proportional torque. This torque acts on the input shaft as well as the output shaft, but the input shaft is fixed. In the control system, the shaft speed n which is produced by a given voltage across the motor is dependent upon the characteristic of the motor and the combined inertia of the motor and load. The mechanical analogy duplicates these characteristics by selection of the load as indicated in Fig. 3.2.

A typical motor characteristic (Fig. 3.2*a*) has the motor torque fall off linearly with increasing speed. The same effect is produced in the mechanical analogy, shown below in Fig. 3.2*a*, by damping vanes *D* which require torque proportional to speed. The inertias *J* in the two systems are equal. The motor characteristic of an a-c motor has very little decrease in torque with speed (Fig. 2.14) and may be approximated as in Fig. 3.2*b*; the mechanical system in this case would have no damping action. A hydraulic valve and piston tends to produce a speed of motion of the shaft independent of pressure, and thus the characteristics may be approximated as shown in Fig. 3.2*c*. Neglecting the effect of inertia compared to damping, the mechanical analogy would have damping vanes only, as shown in the lower part of Fig. 3.2*c*. Other analogies can, of course, be made; in circuits, the same type of response can be

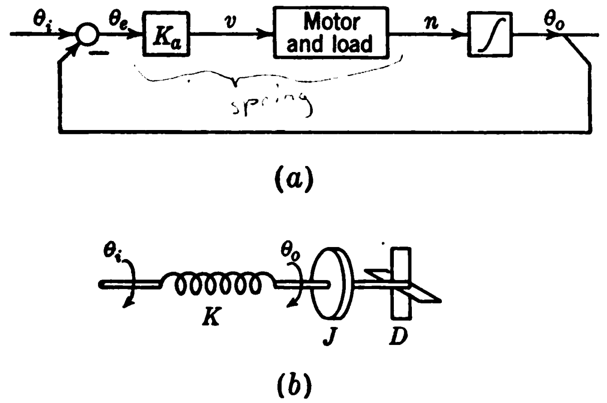


FIG. 3.1. Analogy between control system (*a*) and spring-mass system (*b*).

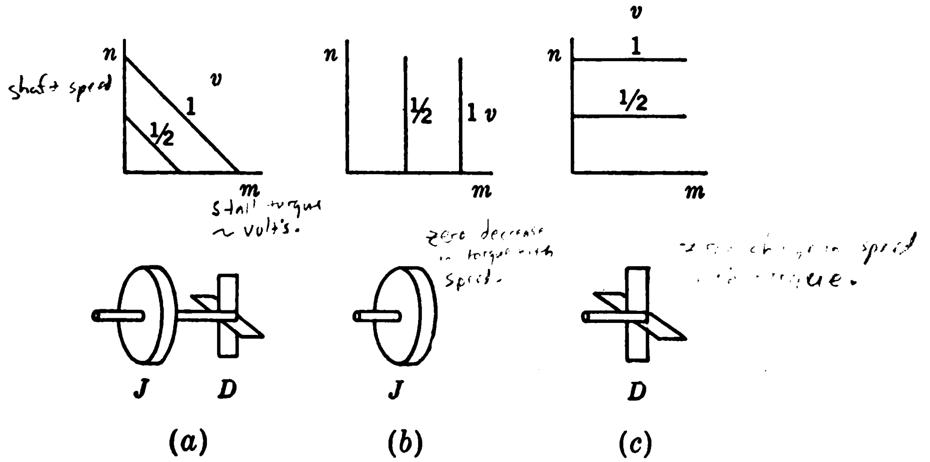

n
shaft speed
v
1
½
m
stall torque ~ volts.
n
½
1 v
m
zero decrease in torque with speed.
v
n
1
½
m
zero change in speed with torque.
J
D
J
D
(a)
(b)
(c)

FIG. 3.2. Analogy between motor characteristics and mechanical load: (*a*) inertia and damping; (*b*) inertia only; (*c*) damping only.

obtained by a simple series circuit involving resistance, inductance, and capacitance. The mechanical analogy, however, has the particular similarity of having the output a shaft rotation just as in the control system. The mechanical system will be used in the remainder of this chapter in the analysis of transients.

3.2. Time Constant

> *A system having a single form of energy storage is described by a first-degree equation. The transient response to a disturbance dies out at a rate which is proportional to the change yet to be made to restore the system to the initial condition. The time which would be required to complete the change if the rate were fixed is a constant called the "time constant."*

Consider the problem of finding the approximate transient response of a sample system direct from physical reasoning (Fig. 3.3). One end of the spring is fixed; the other is twisted through an angle θ_M and then released. The free end has vanes on it which are assumed to require a torque proportional to their angular velocity v. This angular velocity is initially high because the spring torque is high. As the angle of twist decreases, however, the torque decreases, and thus the velocity decreases. For the given assumptions, the velocity of return will always be proportional to the angle of twist remaining. This relation is sufficient to establish the shape of the transient curve shown in Fig. 3.3. Velocity is the slope of this curve; the curve is thus unique in that its slope is proportional to its ordinate. A reasonably accurate curve can be drawn using the above information. The initial slope, for example, can be expressed as

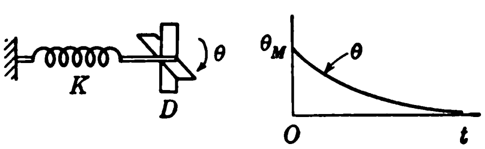


FIG. 3.3. Single-energy-storage system.

$$v_M = \frac{K\theta_M}{D} = \frac{\theta_M}{D/K} \qquad (3.1)$$

A tangent to the initial point of the curve intersects the horizontal axis at D/K, as shown in Fig. 3.4. The curve deviates from this tangent because the torque decreases with decreasing angle, but since velocity is proportional to angle of twist, the ratio between angle of twist and velocity is constant. Thus a tangent drawn to the curve at any point will intersect the horizontal axis at a time D/K farther along the axis from the abscissa of the point. The time constant T is thus

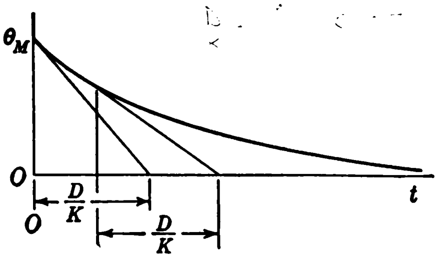


FIG. 3.4. Time constant.

$K\theta = D$ [illegible]

defined as the time in which the angle would reach zero if its velocity were held constant.

Note that the only fact which must be established about a system, to determine that its transient will be of this form, is that the rate of change of the variable is proportional to the change yet to be made. Many control systems, for example, are so set up that, within limitations, the output shaft is driven at a rate proportional to the error angle. In so far as this is achieved, the transient response would be of the same shape as Fig. 3.4. Actually, in either the control system or the mechanical system, the mass of the output element prevents an immediate change in velocity at any time. At any time when the inertia torques are small compared to the damping torques, however, the above analysis applies.

3.3. Exponential Function

The transient response of a first-degree system can be expressed mathematically by an exponential function of time.

The transient response of a first-degree system is shown in Fig. 3.4; the problem here is to determine the equation of this curve. The result, $e^{-t/T}$, is probably familiar to most readers and can be justified immediately if the formula for differentiation is accepted.

$$\frac{d}{dt} e^{-t/T} = -\frac{1}{T} e^{-t/T} \tag{3.2}$$

The derivative is thus proportional to the function itself, as required to match the condition that the slope of the curve is proportional to the ordinate. An attempt to prove this formula by direct use of the definition of differentiation leads to an indeterminate result, however:

$$\frac{d}{dt} e^{-t/T} = \frac{e^{-(t+dt)/t} - e^{-t/T}}{dt} = e^{-t/T} \frac{e^{-dt/T} - 1}{dt} \tag{3.3}$$

The fractional quantity apparently has the limit 0/0 when dt is allowed to approach zero. A separate justification of (3.2) can be made, however, which is suggested by the concepts in the previous topic. Start from the fact that the slope of the curve in Fig. 3.4 is proportional to the ordinate; then one ordinate can be approximated in terms of an adjacent ordinate as

$$x_2 = x_1 - \left[\frac{dx}{dt}\right]_{x=x_1} dt = x_1 + Cx_1\, dt \tag{3.4}$$

The constant C relating slope to ordinate is actually the negative reciprocal of the time constant. The ratio between any two adjacent ordinates which are spaced a fixed interval apart is then constant.

$$\frac{x_2}{x_1} = \frac{x_1(1 + C\,dt)}{x_1} = 1 + C\,dt \tag{3.5}$$

This property of the curve is satisfied by an exponential function of the general form

$$\frac{x_2}{x_1} = \frac{a^{t_1+dt}}{a^{t_1}} = \frac{a^{t_1}a^{dt}}{a^{t_1}} = a^{dt} \tag{3.6}$$

An exponential function thus is capable of fitting the curve in Fig. 3.4. However, this curve is such that its slope is always proportional to its ordinate, which means that an exponential function must have a derivative proportional to its value.

$$\frac{d}{dt}\,a^t = Ca^t \tag{3.7}$$

The base value a may have any value in the above formula; a particularly convenient value, however, would be one which would have the constant $C = 1$. This base proves to be the value $e = 2.71828 \cdots$, as can be shown by the following reasoning: Any function of t can be expressed in terms of its value and derivatives at $t = 0$ by using Maclaurin's series:

$$f(t) = f(0) + f'(0)t + f''(0)\,\frac{t^2}{2} + f'''(0)\,\frac{t^3}{2 \cdot 3} + \cdots \tag{3.8}$$

But if $f(t)$ is equal to e^t, then $f'(t) = e^t$. The successively higher derivatives are also equal to e^t, as indicated by the special case of $f''(t)$:

$$f''(t) = \frac{d}{dt}\,f'(t) = \frac{d}{dt}\,e^t = e^t \tag{3.9}$$

Maclaurin's series for e^t thus becomes

$$e^t = e^0 + e^0t + e^0\,\frac{t^2}{2} + e^0\,\frac{t^3}{2 \cdot 3} + \cdots \tag{3.10}$$

The value of e is evaluated by letting $t = 1$.

$$e = 1 + 1 + \tfrac{1}{2} + \tfrac{1}{6} + \cdots = 2.718 \cdots \tag{3.11}$$

The final step in justifying (3.2) is to introduce a coefficient of t:

$$\frac{d}{dt}\,e^{st} = s\,\frac{de^{st}}{d(st)} = se^{st} \tag{3.12}$$

Note that st is the variable in taking the derivative. This final result proves to be the basis for converting all differential equations to algebraic equations, as will be demonstrated in the next section.

3.4. Classical Method of Solution

An exponential function is assumed and substituted into the differential equation, producing an algebraic equation in which the variable s replaces the operation d/dt.

The differential equation for the angle of twist of the spring of the system in Fig. 3.4 is given by

Transient

$$0 = K\theta + D\frac{d\theta}{dt} \tag{3.13}$$

K = spring const.
D = damping constant

where K is the spring constant and D the damping constant. A solution for θ of the form Ae^{st} is assumed and substituted into this equation.

$$0 = KAe^{st} + DsAe^{st} \tag{3.14}$$

Note that the function Ae^{st} is common to both terms and may be canceled, leaving an algebraic equation in s:

$$0 = K + Ds \tag{3.15}$$

The root of this equation is $s = -K/D$. The assumed solution Ae^{st} thus satisfies the differential equation for the value $s = -K/D$, which satisfies the corresponding algebraic equation. The amplitude A is determined by the initial angle of twist of the spring θ_M at the time that it is released:

$$\theta_M = [Ae^{(-K/D)t}]_{t=0} = A \tag{3.16}$$

This solution $\theta = \theta_M e^{(-K/D)t}$ can be checked against the previous solution in Art. 3.2 by solving for the time which would be required for the angle to reach zero if the rate were held constant:

$$T = \frac{\theta}{-[d\theta/dt]_{t=0}} = \frac{\theta_M e^{(-K/D)t}}{-\theta_M(-K/D)e^{-(K/D)t}} = \frac{D}{K} \tag{3.17}$$

which is the result obtained in Art. 32.

The factor by which the angle of twist is actually decreased during each time-constant interval is

$$\frac{\theta_M e^{-(t+T)/T}}{\theta_M e^{-t/T}} = e^{-T/T} = e^{-1} = 0.37 \tag{3.18}$$

The above example emphasizes the nature of the transient itself; usually many examples arise in which steady-state value of the variable must also be determined. In the system in Fig. 3.1, for example, the input shaft may be suddenly turned through an angle $\theta_i = 1$ rad. The problem is to determine the angle θ_o as a function of time. The differential equation for the case when there is only damping on the output would then be written

$$K\theta_i = K\theta_o + D\frac{d\theta_o}{dt} \tag{3.19}$$

The output angle θ_o is considered to be the sum of a steady-state and a transient response. The steady-state part of θ_o is 1, because the final twist in the shaft after the transient is dissipated is then zero. The transient part is assumed to be $\theta_T = Ae^{st}$ as before. Substituting the sum of these values into (3.19) gives

$$K(1) = K(1 + Ae^{st}) + D(0 + sAe^{st}) \tag{3.20}$$

Separating the terms into steady-state and transient parts, gives

$$K = K + D(0) + Ae^{st}(K + sD) \tag{3.21}$$

The steady-state terms in (3.21) check; *i.e.*, $K = K + 0$. The transient terms merely add zero to the right side of (3.21) if the characteristic function within the parentheses is zero as in (3.15). The assumed form of θ_o is thus correct, and the amplitude A is determined from the condition that θ_o is initially zero:

$$\theta_o = [1 + Ae^{(-K/D)t}]_{t=0} = 0 = 1 + A \tag{3.22}$$

The value of θ_o at the end of one time constant is $1 - e^{-1}$, or 0.63.

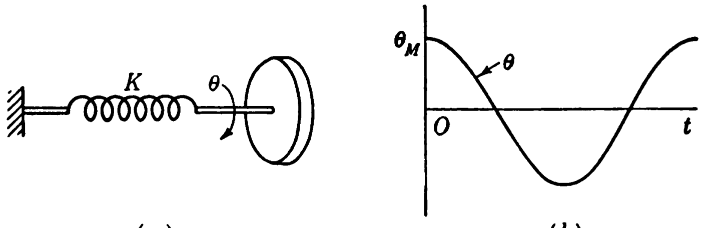
Note that the determination of the form of the transient is the same regardless of the input. Mathematically, the sum of the transient terms in the equation must be zero. Physically, the transients are torques interchanged between elements of the system itself and are independent of input motion.

3.5. Resonant System

The transient response of a spring-mass system is a continuous oscillation, which is sinusoidal in shape. The amplitude is constant because the energy in the system is constant.

Consider the problem of determining the transient response of the system in Fig. 3.5*a* from physical reasoning alone, assuming the damping to be zero. The mass is twisted through an angle of 1 rad and then released at time $t = 0$. The spring torque acting on the inertia causes an acceleration of the inertia proportional to the angle of twist remaining in the spring. The velocity of the inertia thus increases until the angle is zero, meaning that the spring is unstressed. At this time, the momentum of the inertia carries it on until the reversed torque of the spring decelerates the inertia to a stop. The spring must now be twisted through 1 rad in the opposite sense in order to have the same stored energy as when the inertia was first released. The inertia will thus oscillate back and forth, as indicated by the sketch in Fig. 3.5*b*.

As yet, the shape of the curve has not been specified, although the reader probably knows from similar problems that the curve is a sine wave. The purpose of this section is to prove this relation in a way which will be helpful in understanding later representation of such functions by rotating vectors and exponential functions.



(*a*) (*b*)

FIG. 3.5. Transient response of resonant system.

Consider the relationship between angle, velocity, and acceleration of the inertia shown in Fig. 3.6. The acceleration curve must be of the same shape and opposite in sense to the angle curve. An ordinate of the velocity curve varies as the slope of the angle curve, and an ordinate of the acceleration curve varies as the slope of the velocity curve. Performing two identical operations in sequence, starting from the angle curve, gives the acceleration curve of the same shape and suggests that the intermediate curve, that of velocity, may be of the same shape also.

Assume that the shape of the θ curve is sinusoidal and check by differentiation:

$$\theta = \cos \omega t \tag{3.23}$$

$$v = \frac{d\theta}{dt} = (-\omega) \sin \omega t \tag{3.24}$$

$$a = \frac{dv}{dt} = (-\omega)(+\omega) \cos \omega t \tag{3.25}$$

The double differentiation does indeed give an acceleration function which is of the same form as the angle function but opposite in sign. The natural frequency ω is determined by equating spring torque to acceleration torque:

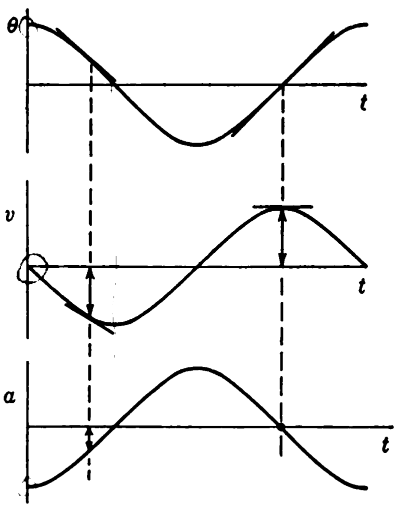


FIG. 3.6. Relationship between angle, velocity, and acceleration curves.

$$K\theta = -Ja = J\omega^2\theta \tag{3.26}$$

or

Spring torque

$$\omega = \sqrt{\frac{K}{J}} \tag{3.27}$$

These results can also be checked by noting that the total energy W in the system is then constant:

Total energy
const

$$W = \tfrac{1}{2}K\theta^2 + \tfrac{1}{2}Jv^2 = \tfrac{1}{2}K[(\theta_M \cos \omega t)^2 + (\theta_M \sin \omega t)^2] \tag{3.28}$$
$$= \tfrac{1}{2}K\theta_M{}^2$$

The next step is to obtain a representation of sine waves which is easier to use.

3.6. Vector Representation of Sine Waves

> *A sine wave is obtained by projecting a rotating vector on a fixed axis. Differentiation shifts a vector ahead* 90°.

A sine wave can be obtained as the projection of a rotating vector on a fixed axis, as shown in Fig. 3.7. The vector has a unit length and is rotating at constant angular velocity ω rad/sec. The angle of this vector with respect to the horizontal axis is ωt, and the projected length of this vector on the vertical axis is $\sin \omega t$.

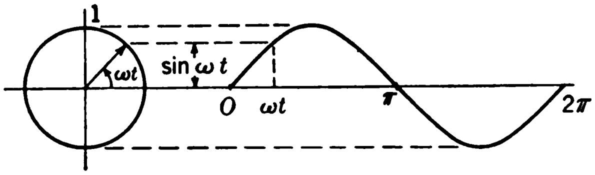
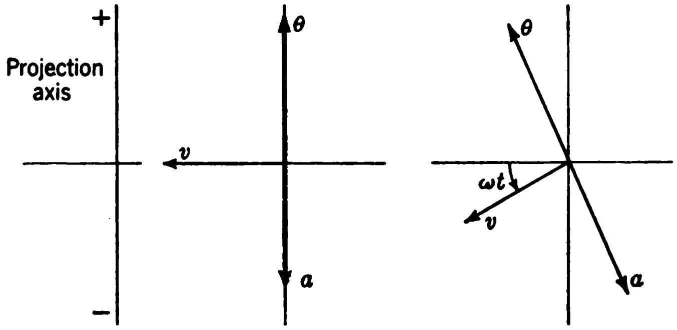


FIG. 3.7. Sine wave as the projection of a rotating vector.



(a) (b)

FIG. 3.8. The variables in Fig. 3.6 as vectors: (a) $\omega t = 0$; (b) $\omega t = 30°$.

Consider now the use of vectors to represent all the sinusoidally varying quantities shown in Fig. 3.6. The initial position of the vectors for $t = 0$ is as shown in Fig. 3.8*a*. Note that the projected length of θ is at a positive maximum, v is zero, and a is at a negative maximum, checking the initial values in Fig. 3.6. At a short interval of time later corresponding to $\omega t = 30°$, all the vectors have rotated through 30° and projected lengths show the same variation as given in Fig. 3.6.

The vector representation is thus very convenient in that a single straight line is sufficient to identify the magnitude and the phase angle of the variable. A convenient notation uses the symbol $A\underline{/\omega t}$ to represent a vector of magnitude A and phase angle ωt. The vector is often considered to be the variable, although the projected length of the vector on a fixed axis represents the actual value of the variable at any instant.

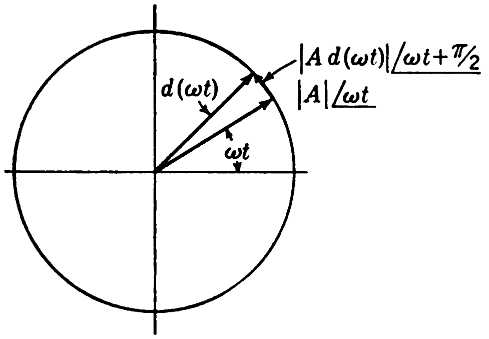

$|A\,d(\omega t)|\underline{/\omega t+\pi/2}$
$|A|\underline{/\omega t}$
$d(\omega t)$
ωt

FIG. 3.9. Derivative of a vector.

An interesting observation is that v, the derivative of θ, is a vector rotated 90° ahead of θ. Similarly, a, the derivative of v, is a vector 90° ahead of v. This suggests that differentiation of a sinusoidal quantity always involves rotating the vector ahead 90°, which can be proved by applying the definition of a derivative to the vector, as shown in Fig. 3.9. A derivative of a function is the limit of the ratio between an increment in the function and the corresponding increment in the variable as the increments approach zero. In this case the vector is a function of time t. Its increment is a short vector connecting two successive positions of the vector as shown. The derivative is then

$$\frac{d}{dt} A\underline{/\omega t} = \frac{A\,d(\omega t)\underline{/\omega t + \pi/2}}{dt} = A\omega\underline{/\omega t + \pi/2} \tag{3.29}$$

The derivative of a vector is thus a vector 90° ahead of the vector with a magnitude ω times that of the vector. Again note that the instantaneous value of the derivative is just the projected length of the vector in Fig. 3.7 on the fixed axis. It is more convenient, however, to carry out the manipulations required in solving a problem in full vector form and obtain real quantities by projecting on a fixed axis as a final step.

3.7. Meaning and Use of the Operator $\sqrt{-1}$

Solving a resonant system in terms of e^{st} *leads to roots of the form* $s = \sqrt{-1}\,\omega$. *Comparison of this result with the vector solution identifies* $\sqrt{-1}$ *as an operator which rotates a vector* 90°.

The transient response of a resonant system is identified as a sine wave in Fig. 3.6 and as a vector in Fig. 3.8. Consider now the concepts which arise in solving the same problem (Fig. 3.5) by assuming a solution of the form e^{st}. The differential equation for the angle of the inertia is

$$0 = K\theta + J\frac{d^2\theta}{dt^2} \tag{3.30}$$

Substitution of the assumed solution $\theta = Ae^{st}$ gives

$$0 = KAe^{st} + Js^2Ae^{st} \tag{3.31}$$

Canceling Ae^{st} leaves the characteristic equation in s:

$$0 = K + s^2J \qquad s^2 = -\frac{K}{J} \tag{3.32}$$

The roots of this equation can be neither a positive nor a negative real number for, in either case, s^2 would be a positive number. One might conclude that a solution of the form e^{st} is impossible for the sinusoidal variation which is known to be the response of this system. Consider, however, the concepts which arise if the solution is carried out in terms of roots of the form

$$s = \pm\sqrt{-1}\sqrt{\frac{K}{J}} = \pm\sqrt{-1}\,\omega$$

The symbol $\sqrt{-1}$ has no meaning assigned to it at this point; it merely permits solution to be completed by expressing the transient as

$$\theta = A_1e^{\sqrt{-1}\omega t} + A_2e^{-\sqrt{-1}\omega t} \tag{3.33}$$

For convenience, neglect the second term at first, and consider the form which a and v must have for the given form of θ.

$$\theta = Ae^{\sqrt{-1}\omega t} \tag{3.34}$$

$$v = \frac{d\theta}{dt} = A\sqrt{-1}\,\omega e^{\sqrt{-1}\omega t} \tag{3.35}$$

$$a = \frac{dv}{dt} = A(\sqrt{-1})^2\omega^2e^{\sqrt{-1}\omega t} \tag{3.36}$$

The differential equation (3.30) is satisfied, as can be checked by substituting for θ from (3.34) and for a from (3.36).

$$0 = KAe^{\sqrt{-1}\omega t} + J(\sqrt{-1})^2\omega^2Ae^{\sqrt{-1}\omega t} \tag{3.37}$$

Substituting $(\sqrt{-1})^2 = -1$ and $\omega^2 = K/J$,

$$0 = Ae^{\sqrt{-1}\omega t}\left(K - \frac{K}{J}J\right) \tag{3.38}$$

The equation is satisfied, with no meaning as yet assigned to $\sqrt{-1}$ or $e^{\sqrt{-1}\omega t}$. Comparison of the form of Eqs. (3.34) to (3.36) with the vectors in Fig. 3.8, however, suggests that $\sqrt{-1}$ must represent rotation of a vector by 90°, as shown in Fig. 3.10. Assuming that $Ae^{\sqrt{-1}\omega t}$ represents the vector for θ, then the vector for v is obtained by rotation through 90°, and the vector for a is obtained by rotation through an additional 90°. The velocity v is then of the same form as obtained by differentiation in (3.29).

To be consistent with the vector diagram in Fig. 3.8*b*, the term $e^{\sqrt{-1}\omega t}$ must correspond to rotation of a vector through the angle ωt, although this will not be checked until the next article. Assuming this to be true, however, the term $e^{-\sqrt{-1}\omega t}$ must represent rotation through the angle $-\omega t$. The complete form of the transient given in (3.33) is then the sum of two oppositely rotating vectors, as shown in Fig. 3.11.

The sum of two terms as in (3.33) must always produce a real number, since the variable θ itself is a real number. It is therefore optional whether the solution is carried out as the sum of oppositely rotating vectors or whether it is considered in terms of the projection of one of the vectors.

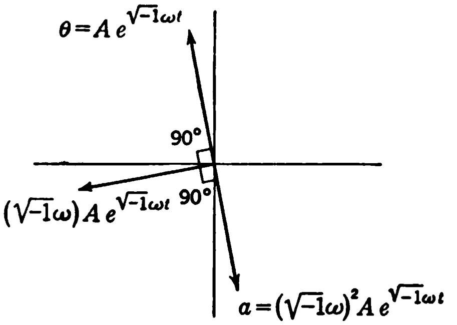

$\theta = Ae^{\sqrt{-1}\omega t}$
90°
$(\sqrt{-1}\omega)Ae^{\sqrt{-1}\omega t}$
90°
$a = (\sqrt{-1}\omega)^2 Ae^{\sqrt{-1}\omega t}$

FIG. 3.10. Meaning of the operator $\sqrt{-1}$.

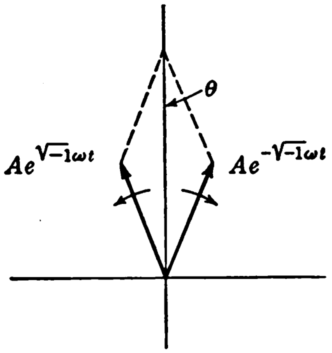


FIG. 3.11. Sum of counter-rotating vectors.

3.8. Calculations with Vectors or Complex Numbers

A vector is related to a complex number by Euler's equation, $e^{j\phi} = \cos\phi + j\sin\phi$. *Addition, multiplication, and differentiation can be carried out in either form, but the vector form is more convenient for the last two operations.*

The comparison of the known vector solution of Art. 3.5 for the transient in a resonant system with that obtained in Art. 3.7 by assuming a solution of the form e^{st} leads to the conclusions that $e^{\sqrt{-1}\omega t}$ must represent rotation of a vector through the angle ωt and that $\sqrt{-1}$ represents rotation through 90°. A check on these conclusions will be made by showing that addition, multiplication, and differentiation of numbers based on the above concepts give consistent and useful results. The reader is probably familiar with these results from previous study of complex numbers, but the following justification of these concepts is believed to be of value.

First, a vector can be expressed as a complex number by using $\sqrt{-1}$, denoted by j, to identify a new axis, as shown in Fig. 3.12. A general vector $A\underline{/\phi}$ is split into two components, one horizontal and the other vertical. The horizontal line to the right of the origin is selected to be the basic reference line. A vertical line is then identified by the factor j, or $\sqrt{-1}$, to mean rotation through 90°. The vector can thus be identified in terms of its components as:

$$A\underline{/\phi} = A \cos \phi + jA \sin \phi \tag{3.39}$$

The addition of two vectors can then be accomplished by the addition of horizontal and vertical components separately:

$$\begin{aligned} A_1\underline{/\phi_1} &+ A_2\underline{/\phi_2} \\ &= (A_1 \cos \phi_1 + jA_1 \sin \phi_1) + (A_2 \cos \phi_2 + jA_2 \sin \phi_2) \\ &= (A_1 \cos \phi_1 + A_2 \cos \phi_2) + j(A_1 \sin \phi_1 + A_2 \sin \phi_2) \end{aligned} \tag{3.40}$$

This result is the same as the familiar addition of vectors by placing the tail of the second vector at the head of the first vector.

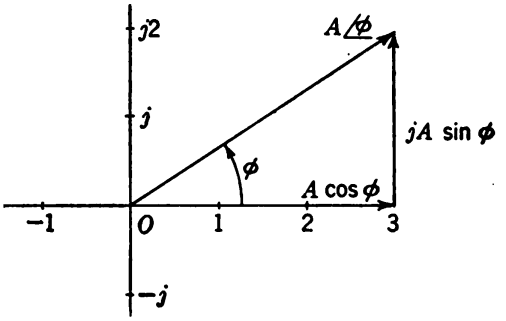


FIG. 3.12. Conversion between vector and complex number.

The process of multiplying two vectors[1] has no significance until it is recalled that a vector at angle ϕ can be represented by $Ae^{j\phi}$. Then multiplication can be carried out using the addition of exponents:

$$A_1e^{j\phi_1}A_2e^{j\phi_2} = A_1A_2e^{j(\phi + \phi_2)} \tag{3.41}$$

Multiplication of vectors must therefore mean multiplication of magnitudes and addition of angles. Check this result by expanding each side of (3.41) in terms of components:

$$\begin{aligned} A_1e^{j\phi_1}A_2e^{j\phi_2} &= A_1(\cos \phi_1 + j \sin \phi_1)A_2(\cos \phi_2 + j \sin \phi_2) \\ &= A_1A_2[\cos \phi_1 \cos \phi_2 + j(\cos \phi_1 \sin \phi_2 + \sin \phi_1 \cos \phi_2) \\ &\qquad + j^2 \sin \phi_1 \sin \phi_2] \end{aligned} \tag{3.42}$$

$$\begin{aligned} A_1A_2e^{j(\phi_1+\phi_2)} &= A_1A_2[(\cos \phi_1 \cos \phi_2 - \sin \phi_1 \sin \phi_2) \\ &\qquad + j(\cos \phi_1 \sin \phi_2 + \sin \phi_1 \cos \phi_2)] \end{aligned} \tag{3.43}$$

The results are seen to be the same, recognizing that j^2 means rotation through 180°, thus changing the sign of the last term in (3.42).

Another check on the meaning of $e^{j\phi}$ can be made by differentiation of each side of the equation

$$e^{j\phi} = \cos \phi + j \sin \phi \tag{3.44}$$

---

[1] The term vector in this text is always used to identify a complex number in terms of its amplitude and angle. The other uses of the term vector to identify a directed magnitude such as a force or displacement have other specific definitions for multiplication which are not involved here.

giving

$$je^{j\phi} = -\sin\phi + j\cos\phi \tag{3.45}$$

The two sides of (3.45) can be compared by a vector diagram, as shown in Fig. 3.13. Differentiation thus produces a rotation of 90°, regardless whether the vector is treated as an exponential function or as a complex number.

Equation (3.44), known as Euler's equation, is basic to operations with complex numbers. A rigorous proof of this equation is based on expressing $e^{j\phi}$ in Maclaurin's series, like that in (3.8). Every other term will contain the factor j. The sum of the terms not involving j will be found to be Maclaurin's series for cos ϕ, and the terms involving j will be found to be Maclaurin's series for sin ϕ.

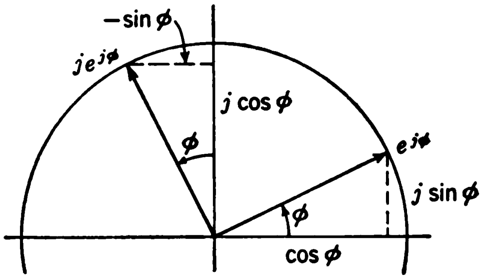


FIG. 3.13. A check on Euler's equation.

A vector, in the sense used in this book, is thus the same as a complex number in that it identifies the position of a point in a plane with respect to a set of axes in that plane. The word vector, however, suggests a line, from the origin to the point, having a magnitude and angle, whereas the word complex number suggests the two orthogonal components. Multiplication and differentiation are the most common operations required in the remainder of the book. These operations are most easily carried out in terms of magnitudes and angles, and hence the word vector is used in preference to complex number.

3.9. System with Damped Oscillation

A solution of the form e^{st} for a system with damped oscillation produces a characteristic equation whose roots $s = \sigma \pm j\omega$ are complex numbers. The real part σ is negative and specifies the damping rate, and ω is the natural frequency in radians per second. Initial conditions are satisfied by use of a vector diagram.

Consider the problem of finding the transient response of the system shown in Fig. 3.14*a*. The addition of damping vanes to the previous system provides a means of dissipating the stored energy in the system so that oscillation will die out, as sketched in Fig. 3.14*b*.

One might assume that the equation of this curve is of the form $\theta = Ae^{\sigma t}\sin(\omega t + \phi)$, because the factor $e^{\sigma t}$ will make the peak amplitudes of the oscillations decay, for σ a negative number, in the same fashion as the first-degree system in Art. 3.4. The amplitude A and

Digitized by Google Original from UNIVERSITY OF MICHIGAN

phase angle ϕ are the constants to be determined by the initial conditions. The differential equation for the input zero is

$$K\theta + D\frac{d\theta}{dt} + J\frac{d^2\theta}{dt^2} = 0 \tag{3.46}$$

A solution can be obtained by substituting $\theta = Ae^{\sigma t} \sin (\omega t + \phi)$, but the algebra becomes tedious. Thus in (3.46) $d^2\theta/dt^2$ will involve four terms. After substituting for θ, the resultant equation can then be split into two equations, stating that the sums of the coefficients of sine terms and cosine terms are zero independently. The first of these equations determines σ, and the second one determines ω. The initial magnitude and rate of change of θ determine A and ϕ. This solution can thus be carried out in terms of real numbers throughout but is impractical for the more complicated systems to be studied later.

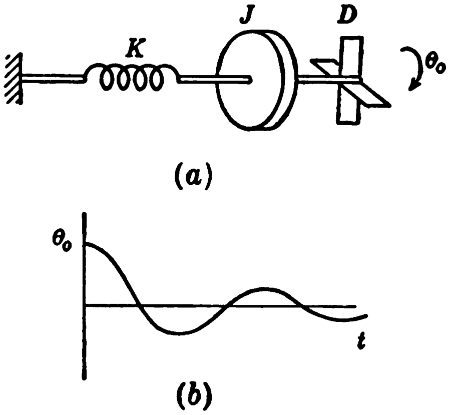


FIG. 3.14. Transient for damped resonant system.

Assume instead a solution of the form Ae^{st} because of the convenience of taking a derivative of such a function. Substituting into the differential equation thus gives

$$KAe^{st} + DsAe^{st} + Js^2Ae^{st} = 0 \tag{3.47}$$

Canceling the common term Ae^{st} leaves the characteristic equation for s:

$$s^2 + \frac{D}{J}s + \frac{K}{J} = 0 \tag{3.48}$$

Solve for the roots by completing the square:

$$s^2 + 2\frac{D}{2J}s + \left(\frac{D}{2J}\right)^2 + \frac{K}{J} - \left(\frac{D}{2J}\right)^2 = 0 \tag{3.49}$$

Factoring the first three terms and transferring the other two to the right side gives

$$\left(s + \frac{D}{2J}\right)^2 = -\left[\frac{K}{J} - \left(\frac{D}{2J}\right)^2\right] \tag{3.50}$$

Solving for s gives complex numbers

$$s = -\frac{D}{2J} \pm j\sqrt{\frac{K}{J} - \left(\frac{D}{2J}\right)^2} \tag{3.51}$$

Denote s for convenience as $\sigma + j\omega$, and note that ω is not the same as that obtained for the resonant system in Art. 3.5. The properties of j

are completely specified in Art. 3.8; all that can or need be done is to complete the solution and interpret the result.

The complex number s is plotted in Fig. 3.15 from (3.51), using known values of K, D, and J. For convenience later, the amplitude is denoted by ω_0 and the angle by ψ. The symbol ω_0 is chosen because this length is $\sqrt{K/J}$, which is the natural frequency for the special case of $D = 0$.

The solution for θ is the sum of two terms, one for each root:

$$\theta = A_1 e^{(\sigma + j\omega)t} + A_2 e^{(\sigma - j\omega)t} = e^{\sigma t}(A_1 e^{j\omega t} + A_2 e^{-j\omega t}) \tag{3.52}$$

But θ is a real quantity; therefore the j parts of the two terms must cancel, leaving twice the real part of either term (Fig. 3.11). Rather than manipulating the algebra, the solution will be completed by treating θ as the projection of a vector $Ae^{\sigma t}e^{j\omega t}$ on a fixed axis (Fig. 3.16). For convenience let the θ vector be horizontal at $t = 0$ and select the axis of projection later.

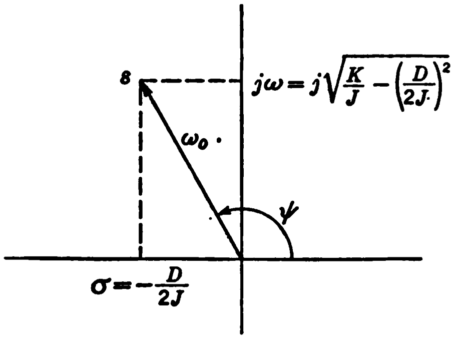

s
$j\omega = j\sqrt{\frac{K}{J} - \left(\frac{D}{2J}\right)^2}$
ω_0
ψ
$\sigma = -\frac{D}{2J}$

Fig. 3.15. Position of complex roots.

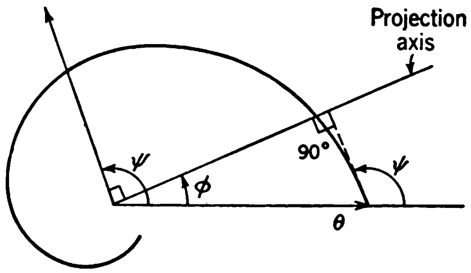


Fig. 3.16. Vector plot for damped resonant system.

The velocity vector v is found by differentiation:

$$v = \frac{d\theta}{dt} = \frac{d}{dt} Ae^{(\sigma + j\omega)t} = A(\sigma + j\omega)e^{(\sigma + j\omega)t} = A\omega_0 e^{j\psi} e^{(\sigma + j\omega)t} \tag{3.53}$$

Thus the vector v is greater than θ by the factor ω_0 and advanced in angle by ψ. As time progresses, both the θ and v vectors decrease in length on account of the factor $e^{\sigma t}$ and rotate counterclockwise at a rate ω rad/sec on account of the factor $e^{j\omega t}$. The tip of the θ vector thus traces a logarithmic spiral curve, as shown.

For the given problem the initial velocity is zero, and therefore the projection axis must be perpendicular to the v vector. The projection of the θ vector at $t = 0$ must be 1, and therefore the length of the θ vector must be $1/\cos\phi$, where ϕ is the angle of the projection axis. The expression for θ for the given initial conditions is therefore

$$\theta = \frac{1}{\cos\phi} e^{\sigma t} \cos(\omega t - \phi) \tag{3.54}$$

3.10. Damping Ratio

> ***The ratio of the negative real part of a root σ to the magnitude of the root ω_0 is a dimensionless quantity, called the damping ratio ζ, which determines the shape of all the characteristic curves of a quadratic system.****

Damped resonant systems are so common that it is well to study their characteristics thoroughly. An important characteristic is the ratio of amplitudes on successive cycles of oscillation. This is given by the factor $e^{\sigma t}$ with the time t being that required by one cycle.

$$e^{\sigma(2\pi/\omega)} = e^{2\pi(\sigma/\omega)} \tag{3.55}$$

In Fig. 3.15 the ratio σ/ω is the cotangent of the angle $180° - \psi$. This ratio varies from zero for $\psi = 90°$ to infinity for $\psi = 180°$. The cosine

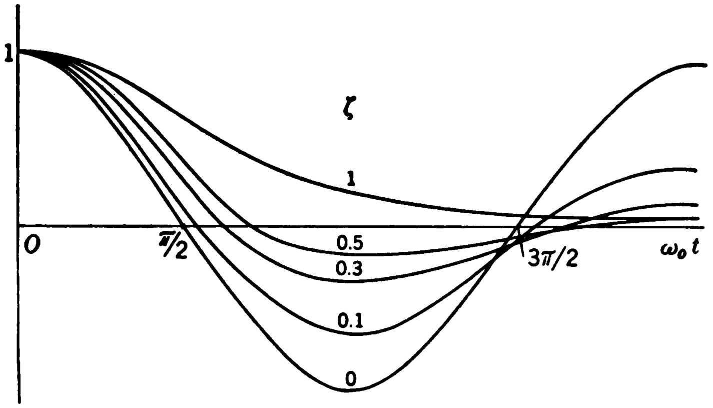


FIG. 3.17. Transients versus damping ratio.

of $180° - \psi$ varies from zero to 1 and is therefore a more convenient quantity to use. The damping ratio is thus defined as

$$\zeta = \cos(180° - \psi) = \frac{\sigma}{\omega_0} \tag{3.56}$$

The decay of the amplitude during one cycle of oscillation, given in (3.55), can then be expressed as

$$e^{2\pi(\sigma/\omega)} = e^{2\pi\sigma/\sqrt{\omega_0^2 - \sigma^2}} = e^{-2\pi/\sqrt{(1/\zeta^2) - 1}} \tag{3.57}$$

The minus sign in the final expression is due to θ being a negative number. The decay per cycle of oscillation thus depends only upon the value of ζ.

A comparison of the transient responses is shown in Fig. 3.17. The curve for a damping ratio ζ of 0 is a continuous oscillation; that for

---

* The general term quadratic system is used to identify any system having two forms of energy storage and thus involving a second-degree characteristic equation.

$\zeta = 1$ is the so-called critical damping case in which the curve just fails to overshoot the zero axis. In control systems for which the damping ratio of the principal pair of complex roots can be adjusted, a value of $\zeta = 0.5$ is often selected because the error reaches zero reasonably fast without excessive overshoot.

3.11. More Complicated Systems

> *The total transient in a linear system of any complexity is the sum of transients of the form e^{st}. The concepts and procedure involved in the solution of such systems can therefore be exactly the same as for the quadratic. Techniques borrowed from Laplace transform theory, however, greatly simplify the algebra.*

The procedure for solving a differential equation developed in this chapter is to substitute an assumed solution of the form ϵ^{st} to obtain an algebraic equation in s. The values of s, called roots, which satisfy this equation correspond to terms of the form $A\epsilon^{st}$ which can exist in the transient response. Specifying the initial conditions determines the amplitudes A of the terms.

In complicated systems it is desirable to describe part of a system by an algebraic quantity called a "transfer function." Close examination of the solution of a differential equation permits such a function to be written by inspection and to be assigned meaning. The Laplace transform theory in turn permits the amplitudes of the transient terms to be written by inspection from the transfer functions without need for restating the initial conditions. These topics are the subject of the next chapter.

CHAPTER 4

MEANING AND USE OF TRANSFER FUNCTIONS

The purpose of this chapter is to develop a rapid method of solving for the response of a system the roots of whose characteristic equation are known. A follow-up system with the feedback loop open is used as an example: the error voltage acts through an amplifier, generator, and then motor to drive the output shaft. Each component, such as the motor, is described by its transfer function, which is an algebraic function of s obtained from the differential equation relating the input and output signals of that component. It will be shown that the steady-state or the transient response of the output shaft in reponse to an error voltage can be written by inspection in terms of the transfer functions. Closing the feedback loop, however, presents a basically new problem in that the forms of the transients are altered.

The Laplace transform method more than fulfills the purpose set forth for this chapter. This method involves "transforming" the differential equation into an algebraic equation by a completely specified process of integration. A set of tables and theorems permits the direct solution for the transient response of any linear system the roots of whose characteristic equation are known. Understanding the concepts involved, however, requires more background than is presented in Chap. 3, and the results are more powerful than needed for the rest of the book. The procedure in this chapter therefore is merely to study closely the pattern of solution used in Chap. 3 and to develop a series of arguments which justify the direct pattern of solution which is permitted by the Laplace transform. A problem is solved by the Laplace transform in Appendix A.

4.1. Layout of Block Diagram

> *A block diagram is a functional layout of an entire system in which the transfer functions of components are shown in blocks and the flow of signals is shown by lines connecting the blocks. Each block is an independent unit in that the connections between blocks do not affect the transfer functions.*

The actual schematic diagram of a simple control system is shown in Fig. 4.1. The difference in angle between the input and output shafts is measured by the potentiometer bridge. The error voltage v_e is amplified electronically to v_f, which is impressed across the field of the generator. The generator voltage v_g in turn is impressed across the motor, which drives the output shaft through an angle θ_o. Note that the line of action of signals is one way: the field voltage determines the generated voltage; the generated voltage does not influence the field voltage. This one-way action permits a simple block diagram to be drawn with the line of action denoted by arrows as in Fig. 4.2.

The change in symbols from v_f to v_1 and from v_g to v_2 denotes a significant change in quantities which must be carefully noted in any analysis. The voltage v_f is the measured voltage on the terminals of the field; the voltage v_1 is the voltage which would exist at the output terminals of the amplifier if the field current were zero. In general, the current flow causes a drop in voltage, making v_f less than v_1. The amplifier determines v_1; the sum R_1 of the amplifier output resistance and the field resistance limits the steady-state value of current. The differential equation for the field also involves the inductance L:

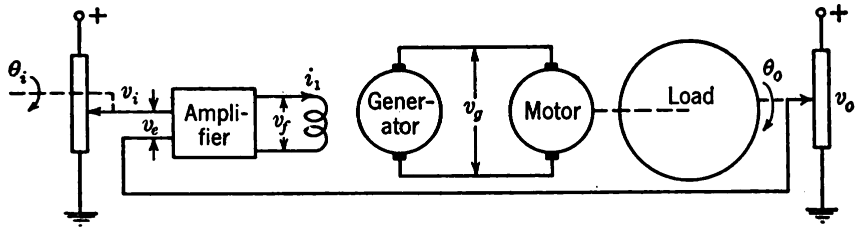


FIG. 4.1. Schematic diagram of a typical servomechanism.

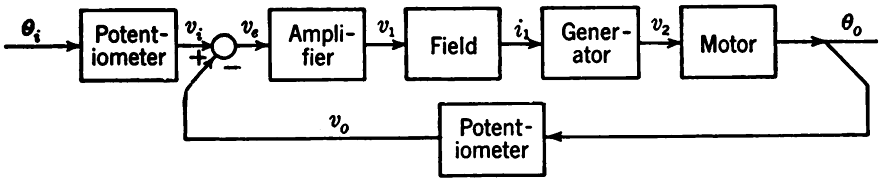


FIG. 4.2. General form of block diagram for Fig. 4.1.

$$v_1 = i_1R_1 + L\frac{di_1}{dt} = R_1\left(i_1 + T_1\frac{di_1}{dt}\right) \tag{4.1}$$

Note that the time constant of the field, T_1, can thus be reduced by having a high output resistance in the amplifier, which can be achieved by using pentodes in the output stage.

The differential equation relating motor speed to impressed voltage on the motor is given in Eq. (2.7). The analysis there is based on the assumption that the voltage impressed across the motor is not changed by the current drawn by the motor. In the given control system, however, the output resistance of the generator, R_g, may be equal to the

Digitized by Google Original from UNIVERSITY OF MICHIGAN

armature resistance of the motor, R_a. The two values are in series, which means that the time constant of the motor, T_2, should be altered by changing R_a to $R_a + R_g$. The speed of the motor must be integrated to determine the angular position of the output shaft.

Other parts of the system, such as the error-measuring potentiometers, are described by a simple constant K_r. The generator voltage per unit field current is K_g.

4.2. The Concept of Transfer Functions

> *A transfer function is obtained by writing the differential equation relating output to input, replacing the operation d/dt by the complex number s, and solving for the ratio of output to input. The results of Chapter 3 will permit the steady-state or transient response of any component to be written by inspection in terms of its transfer function.*

Each component of a control system is described by a differential equation which is rarely higher than second degree. In Chap. 3, the transient response of such components is shown to be expressible in the form e^{st}, in which s is a complex number with a negative real part for damped oscillations. The procedure using the classical method of solution involves substitution of this solution to obtain an algebraic equation in which s appears at each point where the operation d/dt appears in the differential equation. The roots of the algebraic equation are then determined, the transient terms set up in general form, and the amplitudes of the terms found from the initial conditions.

The given problem of finding the output of a component for a given input arises so frequently in the analysis of control systems that it is worthwhile to develop simple rules by which to express this response. These rules will be stated in terms of the transfer function, which is obtained by replacing d/dt in the differential equation by s and solving the resultant equation for the ratio of output to input. Capital letters are used in transfer functions, as noted in the section on terminology. Thus for the generator field the transfer function obtained from (4.1) is

$$\frac{I_1}{V_1}(s) = \frac{1}{R_1(1 + T_1 s)} = \frac{1}{R_1 T_1(1/T_1 + s)} \tag{4.2}$$

Similarly, the transfer function from v_2 to n_2, the motor speed, obtained from (2.7) is

$$\frac{N}{V_2}(s) = \frac{1/K_m}{1 + T_2 s} \tag{4.3}$$

where K_m is the motor constant.

At this point transfer functions are merely reminders of the form of the differential equations. Specific rules will be determined from the solution of the differential equations for two special cases: the steady-state output for an input of the form e^{st}, and the transient response for an input in the form of a short impulse.

Using the concept of transfer functions, the block diagram of Fig. 4.2 may be made definite, as shown in Fig. 4.3. The integration from motor speed to motor angle is noted by the transfer function $1/s$.

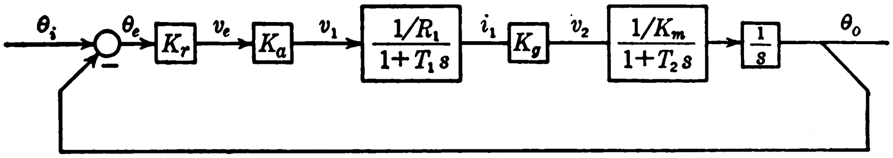
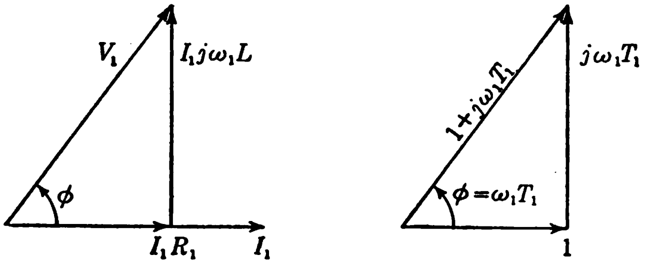


FIG. 4.3. Block diagram using transfer functions.

4.3. Transfer Functions for Exponential Signals

A signal of the form e^{st} acting on a block with a transfer function $F(s)$ is present in the output of that block as $F(s_1)e^{s_1t}$. A definite form of input signal is thus transmitted through a block modified by a factor which is a function of the form of the signal input.

Consider first the example of a sinusoidal voltage being impressed across the generator field. The voltage may be expressed as $V_1e^{j\omega_1t}$.



(a) (b)

FIG. 4.4. Vector diagram for sinusoidal signal.

Assume an output current of the same form, $I_1e^{j\omega_1t}$, and substitute both into the differential equation (4.1).

$$V_1\cancel{e^{j\omega_1t}} = I_1\cancel{e^{j\omega_1t}}R_1(1 + T_1j\omega_1) \tag{4.4}$$

Solve for the ratio of output to input, which is the transfer function for this particular signal.

$$\frac{I_1\cancel{e^{j\omega_1t}}}{V_1\cancel{e^{j\omega_1t}}} = \frac{1}{R_1(1 + T_1j\omega_1)} \tag{4.5}$$

Note that this transfer function could be obtained directly simply by substitution of the particular value $s = j\omega_1$ into the general form (4.2) of the transfer function. The significance of the transfer function is more apparent in the form of a vector than as a complex number, as shown in Fig. 4.4.

The sinusoidal signal is attenuated in magnitude by the factor $1/R_1[1 + (j\omega_1 T_1)^2]$ and shifted in phase by the angle $\tan^{-1} \omega_1 T_1$. The pattern of solution for any differential equation is exactly the same; therefore, the same result applies. The transfer function for sinusoidal signal $e^{j\omega_1 t}$ has the specific value determined by substituting $s = j\omega_1$ into the general form of the function.

Consider now the problem of finding the value of the transfer function for the case of an exponentially decaying signal $Ve^{\sigma_1 t}$, in which σ is negative. The same procedure of assuming a solution of the form $I_1 e^{\sigma_1 t}$ and solving the resulting algebraic equation for the ratio of output to input will give the result

$$\frac{I_1 e^{\sigma_1 t}}{V_1 e^{\sigma_1 t}} = \frac{1}{R_1(1 + T_1\sigma_1)} \tag{4.6}$$

Note that the magnitude of the function $1/1 + T_1\sigma_1$ may be greater than 1, since σ is negative. Physically, the stored inductive energy of the field makes the current present at any time depend upon the voltages previously applied. The applied voltage signal is decaying with time; therefore previous values of voltage cause the current to be larger than V_1/R_1.

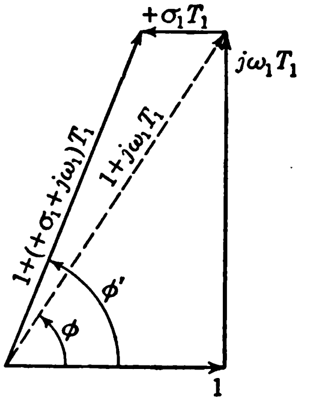


FIG. 4.5. Vector diagram for damped sinusoidal signal.

Finally consider the case of a damped sinusoidal input $V_1 e^{\sigma_1 t} e^{j\omega_1 t}$. The same pattern of solution as used for the sine-wave input will give the same form of result, except that the complex number $\sigma_1 + j\omega_1$ will appear in place of $j\omega_1$. The transfer function will thus be $1/R_1[1 + T_1(\sigma_1 + j\omega_1)]$. The significance of this result is shown in the vector diagram in Fig. 4.5.

Note that the phase shift is greater and the magnitude of the signal is decreased less for a damped sine wave than for an undamped sine wave. This result is used in the root-locus method in Chap. 7.

Thus far analysis of the meaning of transfer functions for applied signals of the form e^{st} has been concerned only with the effect of the component on that signal. In addition to altering the signal, each component may start its own transient, which is the subject of the next section.

4.4. Transient Due to a Unit Impulse

A unit impulse acting on a component with the transfer function $K/(s + 1/T)$ produces a transient of the form $Ke^{s_1 t}$, in which the *amplitude K is obtained from the transfer function by omitting the factor* $s + 1/T$, *and* s_1 *is the value* $s = -1/T$ *which makes this factor zero.*

The transfer function (4.2) of the generator field is obtained from the differential equation (4.1). If the current at $t = 0$ is zero, the impressed voltage v is then equal to $L\,di/dt$. During a short interval Δt, the current will build up to a value

$$I_1 = \int_0^t \frac{di}{dt}\,dt = \int_0^t \frac{v}{L}\,dt = \frac{1}{L}\int_0^t v\,dt \tag{4.7}$$

When the impulse, *i.e.*, the integral of voltage with respect to time, is unity in magnitude, it is called a unit impulse. The approximation in (4.7) is valid as long as the interval Δt is short compared to the time constant T so that iR is small compared to $L\,di/dt$.

The voltage is removed after Δt, and the transient decay of current i can be found using the procedure explained in Art. 3.4. Thus a solution of the form $i = Ie^{st}$ is assumed and substituted into (4.1), giving the characteristic equation $R_1(1 + T_1 s) = 0$. The root of the equation is $s = -1/T_1$; the initial amplitude is given in (4.7), making the response

$$i = Ie^{st} = \frac{1}{R_1 T_1}\,e^{(-1/T_1)t} = \frac{1}{L}\,e^{-(R/L)t} \tag{4.8}$$

This transient is sketched in Fig. 4.6. Note that the transient response to a unit impulse could be written from the final form of the transfer function in (4.2) as Ie^{st}, in which I is the value of the function with the factor $s + 1/T_1$ omitted and s is the value which makes the denominator factor $s + 1/T_1$ equal to 0. The value s which makes the denominator zero is called a pole of the transfer function. For any system with a transfer function having a single pole the transient response for a unit impulse could be written directly using the above results without need for repeating the complete procedure.

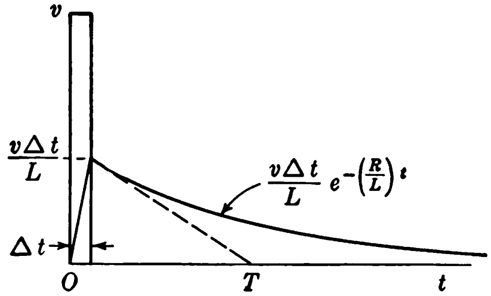


FIG. 4.6. Transient response to unit impulse.

The special case of an impulse acting on the transfer function $1/s$ gives the resultant transient e^{0t} or simply 1. The function $1/s$ arises from the differential equation stating that the speed of the motor shaft is the derivative of the shaft angle. The result is general, however, that the output of a device with transfer function $1/s$ for a unit-impulse input is a unit step. If a unit step were the input to a transfer function of the form $1/(1 + Ts)$, the final output can be shown to

$$\theta = 1 - e^{-t/T} \tag{4.9}$$

An interesting question arises: Can the sequence in which the signal passes through the two functions be interchanged without changing the final output? If the impulse acts on the transfer function $1/(1 + Ts)$ first, the transient will be $1/Te^{-t/T}$. Integrating this signal then gives

$$\theta = \int_0^t \frac{1}{T} e^{-t/T}\,dt = [-e^{-t/T}]_0^t = 1 - e^{-t/T} \tag{4.10}$$

The result in this case is not changed by the order of the operations. The possibility of the interchange of any sequence of blocks will therefore be investigated.

4.5. Interchangeability of Blocks

> *A system having a number of blocks in sequence can have the order of the blocks interchanged without changing the final output signal for a given input signal. This result is true for transients developed within the system as well as for the input signal itself.*

The interchangeability of blocks without changing the output for a given input may seem to be perfectly reasonable but could hardly be said to be obvious. The result, and the concepts used in justifying it, are important in that the justification of the formula for the amplitude of terms will depend upon it.

First assume that a sinusoidal voltage is impressed at v_e in Fig. 4.3 with the feedback loop broken. The effect of the field transfer function on this signal is to reduce the magnitude and shift the phase angle of the current by $1/R_1(1 + T_1 j\omega)$. Similarly, the sinusoidal voltage across the motor produces an output speed which is shifted in phase angle and attenuated in magnitude by $1/K_m(1 + T_2 j\omega)$. The total phase-angle shift is the sum of the shifts in angles, and the attenuation is the product of the attenuations. If the two transfer functions were interchanged in their position in the block diagram, the total angle would be the same, and the attenuation would be the same; therefore the order of the blocks is not significant for a sinusoidal signal. The same type of argument will show that the order of the blocks is not significant for any signal of the form e^{st}. The same argument will of course apply to any type of transfer function in the blocks because the function will be a complex number having a magnitude and phase angle.

The reasoning concerning transients developed in the system is not so direct. Any impulse acting on the system will in general start a transient at each block containing an energy-storage element. Restating this in terms of transfer functions, a transient of the form e^{st} will be set up for each pole of a transfer function. When several poles have the same value, the transient is a special case, treated in Art. 11.3. A unit impulse at $t = 0$ will therefore produce an output of the general form

$$\theta_o = A_1e^{s_1t} + A_2e^{s_2t} + A_3e^{s_3t} + \cdots \tag{4.11}$$

The amplitudes of the terms have not yet been specified. Each of these terms will be present, however, regardless of the order of the blocks in the sequence. One method of determining the amplitudes is to set up the simultaneous equations for the initial conditions of the output angle and solve for the amplitudes. The initial conditions are independent of the sequence of the blocks, however, and therefore the amplitudes of the terms must also be independent of the sequence. The same argument holds for any number of blocks in sequence; the number of transient terms and the number of initial conditions merely increase. Moreover, if the argument is valid for a single impulse, then it is valid for any input, because any input is a series of impulses.

A sequence of blocks may therefore have the transfer functions arranged in any order without changing the final output for a given input. This result is based on the linearity of the system and the definition of a transfer function. Thus if the amplitudes of the signals are so large that the question of a component reaching a nonlinear range is involved, then the original sequence must be preserved. Recall also from Art. 4.1 that a transfer function describes a part of the system selected so that its output signal is not affected by the component represented in the next transfer function.

4.6. Amplitude of Transient Terms

An impulse input to a sequence of blocks produces an output which is the sum of the transients generated by each transfer function. The amplitude of any transient can be found by considering the transfer function which produces it to be the first in the sequence.

Consider the system shown in Fig. 4.3, with the feedback loop still open. A unit impulse is impressed at θ_e; the output angle θ_o as a function of time is desired. The known facts about the system are the response of a transfer function to a signal of the form e^{st}, the transient response to a unit impulse, and the interchangeability of blocks. These facts are sufficient to solve the problem at hand, as can be shown by the following reasoning:

The output angle may be expressed as the sum of the transients arising from each block:

$$\theta_o = A_1e^{s_1t} + A_2e^{s_2t} + A_3e^{s_3t}$$

The unit impulse at θ_e is multiplied first by K_r and K_a in passing through the potentiometer and amplifier. This impulse acting on the field then produces an output $(K_aK_r/R_1T_1)e^{(-1/T_1)t}$, as developed in Art. 4.4. This signal will then be transmitted through the remaining blocks as specified by substituting $s = -1/T_1$ in the transfer functions. The transient term due to the field is thus

$$A_1e^{s_1t} = \frac{K_rK_a}{R_1T_1}e^{(-1/T_1)t}\frac{K_g/K_m}{1 + T_2(-1/T_1)}\frac{1}{-1/T_1}$$

The term $A_2e^{s_2t}$ due to the motor will be found by moving the transfer function associated with the motor to the front of the sequence, as shown in Fig. 4.7. The unit impulse acting on this block then produces the output $(1/K_mT_2)e^{(-1/T_2)t}$, as sketched below the output of the first block. The specific value $s = -1/T_2$ is then substituted in the remaining blocks, with the final output noted in Fig. 4.7. Finally the term $A_3e^{s_3t}$ associated with the pole $s = 0$ is obtained by shifting the block $1/s$ to the front. Substituting $s = 0$ into the other transfer functions makes this term simply $K_rK_aK_g/R_1K_m$. The complete transient response for θ_o for a unit impulse at θ_e in Fig. 4.3, with the feedback loop open, is thus

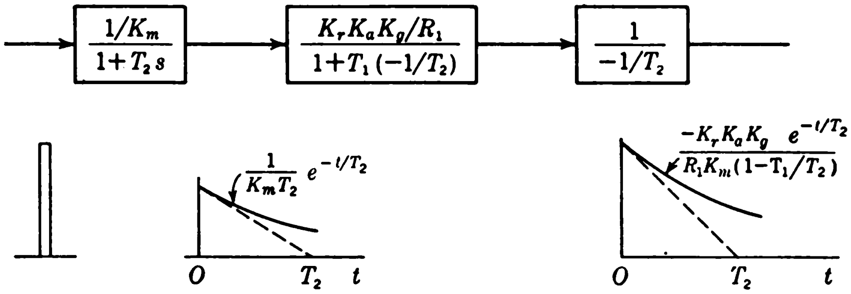

$\frac{1/K_m}{1+T_2s}$
$\frac{K_rK_aK_g/R_1}{1+T_1(-1/T_2)}$
$\frac{1}{-1/T_2}$
$\frac{1}{K_mT_2}e^{-t/T_2}$
O
T_2
t
$\frac{-K_rK_aK_g}{R_1K_m(1-T_1/T_2)}e^{-t/T_2}$
O
T_2
t

FIG. 4.7. Amplitude of a transient term.

$$\theta_o = \frac{K_rK_aK_g}{R_1K_m}\left\{1 + \frac{(1/T_1)e^{-t/T_1}}{[1 + T_2(-1/T_1)](-1/T_1)} + \frac{(1/T_2)e^{-t/T_2}}{[1 + T_1(-1/T_2)](-1/T_2)}\right\} \tag{4.12}$$

This same procedure can apply directly to a more complicated system in which more poles are involved. The factors of the denominator must be known, however, because the poles of the transfer functions determine the terms of the transient.

Additional complications sometimes arise, such as two equal time constants appearing in the system. These are special cases which add nothing to the concepts being considered here; these cases are considered in Chap. 11. The case when a function of s appears in the numerator of a transfer function, however, is important. A simple example will be used to justify the method of handling any such function. If the output-to-input ratio X_o/X_i is $1 + Ts$, then the corresponding differential equation must be

$$x_o = x_i + T\frac{dx_i}{dt} \tag{4.13}$$

Physically this means that the output of this block involves the derivative of the input as well as the input itself. No energy storage is involved, and therefore no transient response is involved. Its transfer effect upon input signals can be readily determined. Assume that $x_i = e^{s_1t}$; then

$$x_o = e^{s_1t} + Ts_1e^{s_1t} = (1 + Ts_1)e^{s_1t} \tag{4.14}$$

A numerator transfer function thus modifies a signal of the form e^{s_1t} as specified by substituting $s = s_1$ into the function. No transient terms arise.

4.7. Exponential Inputs and Initial Conditions

> *An exponential signal into a system can be replaced by an impulse acting through a transfer function which will produce that signal as its output. Initial conditions may be inserted into a system by impulse inputs at proper points in the block diagram.*

The response of a system to a unit impulse is so easy to express that it is worthwhile to build upon this fact if possible. Thus consider the case in which the error voltage impressed at the input to the amplifier is $e^{-\alpha t}$. Note that this is exactly the same form of a function of time which could have been produced by a simple transfer function ahead of the amplifier. One might guess at the transfer function needed in such a preceding block in order that a unit impulse acting on it would produce the given function of time. This is essentially a trial-and-error process in which the rule for the transient produced by an impulse provides the check. Thus the first guess for a transfer function might well be $1/(1 + s/\alpha)$; checking the function of the time produced by a unit impulse gives the results $\alpha e^{-\alpha t}$, as shown in Art. 4.4. The transfer function must therefore be $1/(s + \alpha)$.

Suppose that the input voltage to the amplifier is $\cos \omega t$. What is the transfer function needed to produce this signal from a unit impulse? First recall that a cosine function can be represented by two vectors rotating in opposite directions, as shown in Fig. 4.8*a*. The response $\frac{1}{2}e^{j\omega t}$ can be produced by the transfer function $1/[2(s - j\omega)]$ and the vector $\frac{1}{2}e^{-j\omega t}$ by the transfer function $1/[2(s + j\omega)]$. The block diagram to produce the cosine function from a unit impulse is thus as shown in

Fig. 4.8*b*. The two parallel paths can be replaced by a single block by direct addition of the transfer functions:

$$\frac{1}{2(s - j\omega)} + \frac{1}{2(s + j\omega)} = \frac{s}{(s - j\omega)(s + j\omega)} = \frac{s}{s^2 + \omega^2} \tag{4.15}$$

These examples serve to illustrate the method of attack in guessing the transfer function needed to produce a given input signal from a unit impulse.

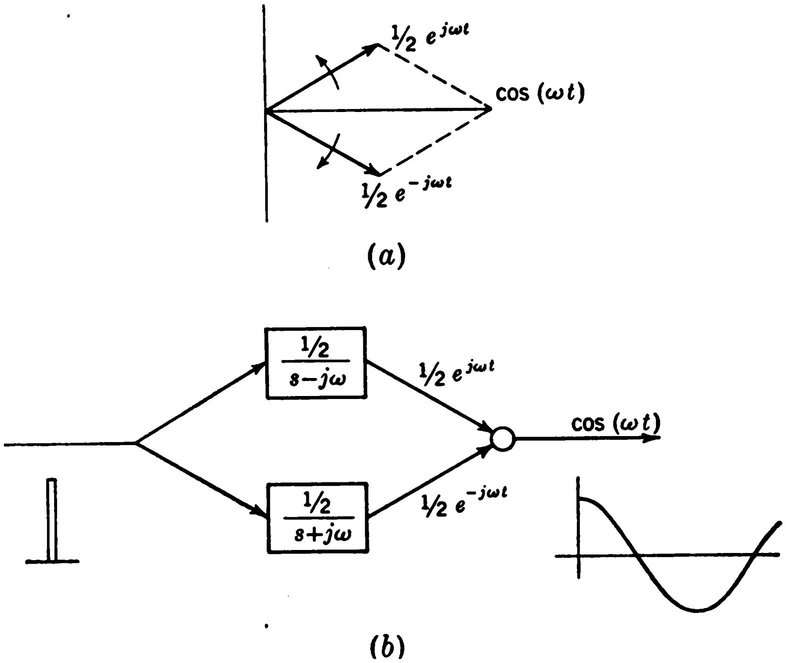

$\frac{1}{2} e^{j\omega t}$
cos (ωt)
$\frac{1}{2} e^{-j\omega t}$
(*a*)
$\frac{1/2}{s - j\omega}$
$\frac{1}{2} e^{j\omega t}$
cos (ωt)
$\frac{1/2}{s + j\omega}$
$\frac{1}{2} e^{-j\omega t}$
(*b*)

Fig. 4.8. Development of cosine wave: (*a*) counterrotating vectors; (*b*) parallel paths.

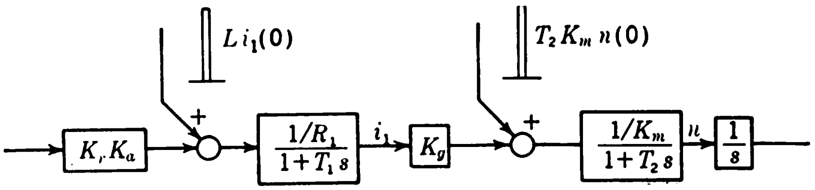


Fig. 4.9. Addition of initial conditions.

Consider now the problem of including initial conditions in an analysis. Suppose, for example, that the field current has an initial value $i_1(0)$. Actually such an initial value would exist because of voltage impressed on the field prior to the time which is arbitrarily selected to be $t = 0$. The response after the time $t = 0$ can be duplicated, however, if a single voltage impulse is impressed on the field at the time $t = 0$ sufficient to bring the current to its specified initial value. In this case the impulse would be $Li_1(0)$, as shown in the block diagram in Fig. 4.9.

Note that an initial speed $n(0)$ can also be included by impressing a voltage impulse $T_2K_mn(0)$ across the motor.

4.8. Feedback as Superposition of Signals

Closing a feedback loop around a block makes the input to the block a function of its own output. An attempt to solve this by superposition of signals leads to a geometric progression. The summation formula is a function of s for which the poles must be found before the transient response can be expressed.

The analysis thus far has developed a complete solution of the original system shown in Fig. 4.3 except that the feedback loop has not yet been closed. The output of the block for a known input can be written by inspection; therefore a solution for the closed-loop system based on the open-loop response appears to be desirable. Such a solution can be expressed by considering each output signal to be a new input signal to the block, as indicated in Fig. 4.10.

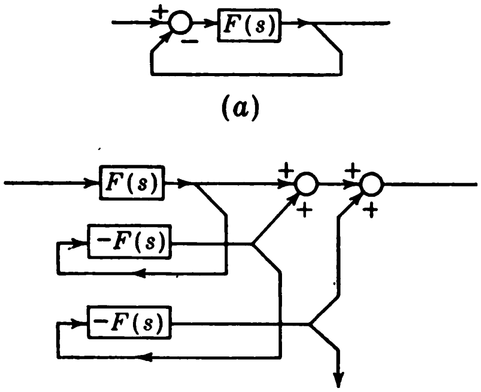


Fig. 4.10. Superposition of feedback signals.

A unit-impulse input to the block having a transfer function $F(s)$ produces an output. This output is in turn fed back through the block producing an additional output. This process continues, with each new portion of output signal being fed back to produce still another part of the output signal. The total output is thus a sum of terms as indicated. The transfer function from the input to the total output is then of the form:

$$\frac{\Theta_o}{\Theta_i}(s) = F(s) - [F(s)]^2 + [F(s)]^3 - [F(s)]^4 + \cdots \tag{4.16}$$

An approximation to the output can be made by considering only the first few terms of the series. The approximation will be reasonably good for the first portion of the transient, but continuing the series soon becomes impractical. A simpler expression of the over-all transfer function may be obtained by recognizing that the series is a geometric progression with the ratio $-F(s)$, for which the summation formula is

$$\frac{\Theta_o}{\Theta_i}(s) = \frac{F(s)}{1 + F(s)} \tag{4.17}$$

The poles of this transfer function are the zeros of the function $1 + F(s)$. These values of s correspond to transients of the form e^{st} which exist for the closed-loop system. The same relationship between θ_i and θ_o may be found more directly as shown below:

$$\frac{\Theta_i}{\Theta_o}(s) = 1 + \frac{\Theta_e}{\Theta_o}(s) = 1 + \frac{1}{F(s)} = \frac{1 + F(s)}{F(s)} \tag{4.18}$$

The zeros of (4.18) will thus be recognized as being the roots of the characteristic equation.

Thus no matter what the approach might be, the roots of the characteristic equation must be found in order to express the transient response as a sum of terms of the form e^{st}.

4.9. Summary

If the transfer function from an input θ_i to an output θ_o is known in factored form, then the transient response of θ_o for a unit impulse at θ_i can be written by simple rules summarized in this article.

Given the transfer function from θ_i to θ_o in factored form:

$$\frac{\Theta_o}{\Theta_i}(s) = \frac{(1 - s/q_1)(1 - s/q_2)\cdots}{(1 - s/p_1)(1 - s/p_2)\cdots} \tag{4.19}$$

The input θ_i is a unit impulse; the output transient is then

$$\theta_o = A_1e^{p_1t} + A_2e^{p\,t} + \cdots \tag{4.20}$$

in which

$$A_1 = \left[(s - p_1)\frac{\Theta_o}{\Theta_i}(s)\right]_{s=p_1} \tag{4.21}$$

In Laplace transform terminology, the function (4.19) from which the output transient is computed is called the transform of θ_o and is denoted as $\Theta_o(s)$. The function of s from which θ_i can be found is in turn called the transform of θ_i and is denoted as $\Theta_i(s)$. The interchangeability of transfer functions demonstrated in Art. 4.5 in turn justifies the relation

$$\Theta_o(s) = \frac{\Theta_o}{\Theta_i}(s)\Theta_i(s) \tag{4.22}$$

The Laplace transform itself provides a rigorous basis for the relationships developed in this chapter: Appendix A contains a problem solved by the Laplace transform method and a concept of the meaning of a transform.

CHAPTER 5

THE FREQUENCY-RESPONSE METHOD

The control of the output of a system by a signal which depends upon the error between input and output is the basic reason for the accuracy of a control system. When an error signal is amplified sufficiently, however, the system will almost always develop a continuous oscillation with no external input being required. One of the prime advantages of the frequency-response method is that it provides some understanding of the requirements for this condition of instability and suggests means of avoiding it. An oscillation of the output when the input is fixed means that the error signal must be simply the output signal itself, changed in sign. The error signal, however, is transmitted through the amplifier and actuator components of the system to produce this output. This transmission must, therefore, involve shifting the error signal by 180° and keeping the net amplitude ratio unity. The frequency-response method uses the ratio of output to error for sinusoidal signals of various frequencies to determine the changes necessary to avoid instability. The transient response of the system is then determined separately or approximated from the frequency-response characteristic.

5.1. Difficulties with the Classical Method

The classical method involves setting up the characteristic equation of the system in polynomial form, inserting numerical values, and solving for the roots. This particular procedure, however, loses the identity of the physical parameters of the system and thus does not suggest the changes needed to improve the system.

A simple cubic system (Fig. 5.1) will serve to illustrate the classical method of solution, although a higher-order system presents the real difficulties. The differential equation relating θ_e to θ_o, represented by the transfer function in the block, is

$$K\theta_e = \left(1 + T_1 \frac{d}{dt}\right)\left(1 + T_2 \frac{d}{dt}\right)\frac{d\theta_o}{dt} \tag{5.1}$$

The equation represented by the junction point is

$$\theta_i = \theta_o + \theta_e \tag{5.2}$$

Substitution from (5.1) into (5.2) gives

$$\theta_i = \theta_o + \left[\frac{1}{K}\left(1 + T_1\frac{d}{dt}\right)\left(1 + T_2\frac{d}{dt}\right)\right]\frac{d\theta_o}{dt} \tag{5.3}$$

The transient response of the system consists of functions of the form Ae^{st} which will satisfy the equation when the input θ_i is zero. Substituting $\theta_o = Ae^{st}$ into (5.3),

$$0 = Ae^{st} + \left[\frac{1}{K}(1 + T_1 s)(1 + T_2 s)\right] sAe^{st} \tag{5.4}$$

Note that the factor Ae^{st} is common to all terms and may be canceled, leaving an algebraic equation in s:

$$0 = 1 + \frac{1}{K}(1 + T_1 s)(1 + T_2 s)s \tag{5.5}$$

This equation is called the "characteristic equation" of the system, and the values of s which satisfy it are called "roots." Each root, such as s, corresponds to a function of the form Ae^{st} which will satisfy (5.3) and thus can exist as a transient in the system.

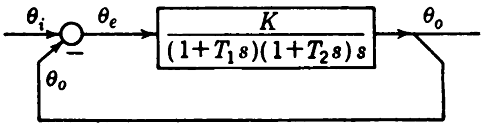


Fig. 5.1. Block diagram for cubic system.

The classic procedure at this point is to expand the function in (5.5) to a polynomial of the form

$$a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = 0 \tag{5.6}$$

In this case the result is

$$\frac{T_1T_2}{K}s^3 + \frac{T_1 + T_2}{K}s^2 + \frac{s}{K} + 1 = 0 \tag{5.7}$$

Note that the coefficients of this polynomial are combinations of the original system parameters T_1, T_2, and K. The addition of another function $1/(1 + T_3 s)$ in the transfer function from θ_e to θ_o would raise the equation to fourth degree, and the complexity of the coefficients would become greater. Note that this complexity arises even though the new time constant T_3 may be short compared to the other ones, T_1 and T_2, and therefore might be reasonably expected to have only a small effect on the roots of the equation.

Several different procedures to find the roots of an equation of this form might be used. One of the simplest procedures is Routh's criterion (see Appendix D) to determine whether any root has a positive real part. Such a root would correspond to a transient which would build up with time, and a system with such a root would be unstable. Consider a third-degree system in which numerical values for the parameters are substituted and the coefficients calculated:

$$a_3s^3 + a_2s^2 + a_1s + a_0 = 0 \tag{5.8}$$

The system can be shown to be stable if

$$\boxed{a_2a_1 - a_3a_0 > 0} \tag{5.9}$$

A more complete statement and a simple proof of this criterion are given in Appendix D. The advantage of understanding the proof is that simplifications in the calculations are then permissible, and the criterion can be extended to determine whether any root has less than a specified damping constant or damping ratio.

Various extensions of Routh's criterion have been devised; these usually involve plotting the limits of stability in terms of system parameters. Regardless of the merits or possibilities of these procedures, the frequency-response method is now much more widely used because its concepts and plots suggest means of improving the system.

5.2. Frequency-response Concepts

> *The concepts developed by the frequency-response method are some of its most valuable contributions to the study of control systems. These concepts will, therefore, be considered first, independent of the calculation details.*

The simplicity of the frequency-response method is achieved by breaking the over-all problem into several small steps. First a component is described by the transmission of a single sine wave. If a test is made, the measurements are not made until all transients have died out. The component is then described by the phase shift between the input and output sine waves and the ratio of magnitudes of these waves. For two components in series, the output of the first component is the input to the second. The total phase shift of a sine wave is then the sum of the individual phase shifts; the amplitude ratio is the product of the individual amplitude ratios. The entire part of the system from θ_e to θ_o in Fig. 5.1 is treated as a unit in Fig. 5.2.

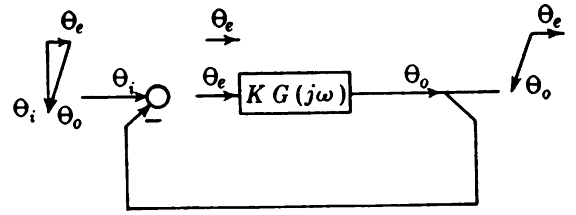


FIG. 5.2. Vector representation of sinusoidal signals at various points around the control loop.

The sine waves are all represented by the corresponding vectors;

instantaneous values can be obtained by projecting the vectors onto a fixed axis. The Θ_e vector is chosen as the reference and has unit magnitude and zero angle. The loop is broken at Θ_e, but the input which would be necessary to produce the given Θ_e vector can be found by a simple vector diagram, as shown on the left side of Fig. 5.2. The above process could be carried out step by step for a very complicated system in order to find the relative magnitudes and phase angles between sine waves which must exist in the steady state.

The vector picture described above for one frequency could be obtained for any other frequency. Keeping Θ_e as a reference, the locus of the Θ_o vector for the various frequencies noted in parentheses is shown in Fig. 5.3.

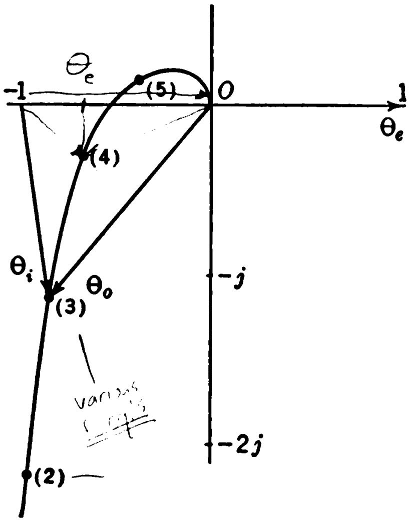


FIG. 5.3. Output and input vectors for fixed error signals of various frequencies.

Note that the head of the Θ_i vector follows the same locus. The ratio of magnitudes of the Θ_o vector to the Θ_i vector is seen to be greatest at about $\omega = 4$. The system will be shown to be stable for this shape of locus if the intersection of the locus with the negative real axis is inside the -1 point.

These concepts are independent of the details of the various plots used in the frequency-response method. The transfer function of a block for a given frequency, for example, might be measured directly from the component itself or calculated from the equations. The plot versus frequency can be made as a vector locus or as separate plots of magnitude and phase angle. In the latter case logarithmic coordinates prove to be convenient. These details, though necessary in calculations for a given system, should not be allowed to confuse the over-all picture.

5.3. Single-time-delay Components

The most common form of transfer function which arises in control systems is that of a simple time delay, $+1/(1 + Ts)$. *Its response with increasing frequency will be shown to start at unity and to develop phase lag and attenuation approaching* 90° *lag and complete attenuation.*

The transfer function $1/(1 + Ts)$ arises from the differential equation

$$\theta_i = (1 + Ts)\theta_o = \theta_o + T\frac{d\theta_o}{dt} \tag{5.10}$$

This form of equation arises for a component having a single form of energy storage, such as the inductive field and motor described in Chap. 2. The relationship between θ_i and θ_o is desired; therefore θ_o will be taken as the reference, where θ_o is equal to $\Theta_o \sin \omega t$ or $\Theta_o e^{j\omega t}$. The vector Θ_o is shown in Fig. 5.4, with the instantaneous value being the projection on the vertical axis. The derivative of θ_o is a sine wave 90° ahead of θ_o, as shown by

$$T\frac{d(\Theta_o e^{j\omega t})}{dt} = T\Theta_o j\omega e^{j\omega t} \tag{5.11}$$

The vector for Θ_i is then the vector sum of Θ_o and $j\omega T\Theta_o$. The locus of Θ_i for a given Θ_o and various values of ω is a vertical line passing through the head of the Θ_o vector. The ratio of input to output of a component is called the *inverse transfer function* because the ratio of output to input is usually considered. The direct transfer function can be plotted by fixing Θ_i as the reference vector and keeping the relationship between vectors the same as in Fig. 5.4. The result is that Θ_o traces a semicircle as ω varies because the angle between the Θ_o vector and the $j\omega T\Theta_o$ vector must be 90° (see Fig. 5.5).

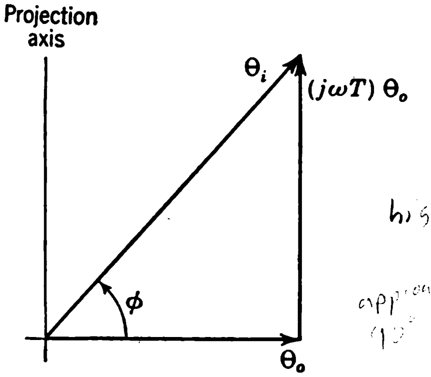


FIG. 5.4. Voltage required to produce a sinusoidal current in an inductive field.

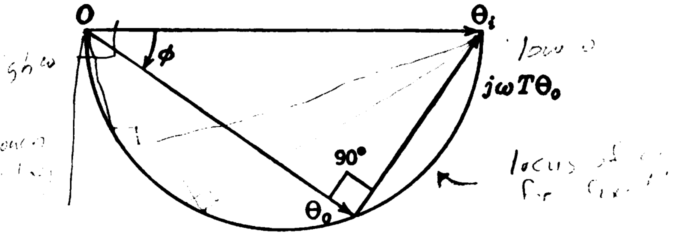
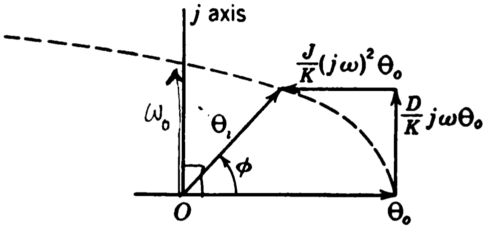


FIG. 5.5. Locus of output vector of component with transfer function $1/(1 + Ts)$ for fixed input vector.

5.4. Quadratic Components

Components having two forms of energy storage have a transfer function of quadratic form. If the damping is greater than a critical value, the quadratic can be factored into two simple time-delay functions. If the damping is considerably less than critical, the system is resonant at a particular frequency.

A mechanical system with a spring and mass or an electrical system with inductance and capacitance have two forms of energy storage. The torsion system shown in Fig. 3.14*a* is used as an example in this article, except that the left end of the spring is driven through an input angle θ_i. The differential equation is then



j axis
$\frac{J}{K}(j\omega)^2\Theta_o$
$\frac{D}{K}j\omega\Theta_o$
ω_0
Θ_i
ϕ
O
Θ_o

FIG. 5.6. Input required to produce fixed output for various frequencies.

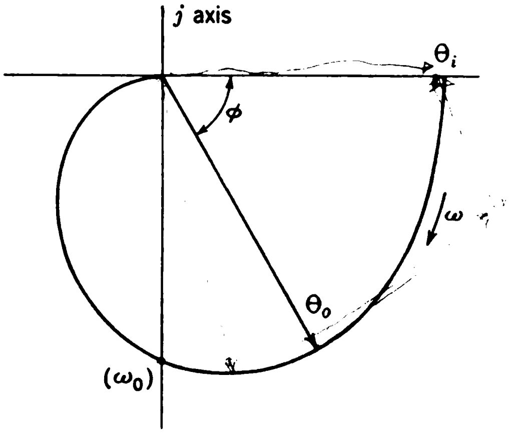


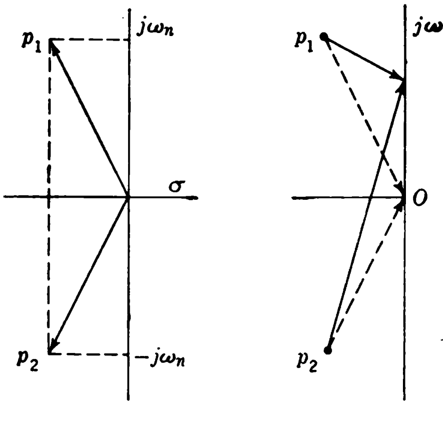
FIG. 5.7. Locus of output for fixed input of various frequencies.

$$K(\theta_i - \theta_o) = D\frac{d\theta_o}{dt} + J\frac{d^2\theta_o}{dt^2} \tag{5.12}$$

For studying frequency response let $\theta_i = \Theta_i e^{j\omega t}$ and $\theta_o = \Theta_o e^{j\omega t}$ in (5.12). Simplifying the resulting equation gives

$$\Theta_i = \Theta_o + \frac{D}{K}(j\omega)\Theta_o + \frac{J}{K}(j\omega)^2\Theta_o \tag{5.13}$$

The output angle Θ_o is the reference vector; the dotted line of Fig. 5.6 is the locus of the Θ_i vector necessary for various values of ω. ~~Note that Θ_i is initially~~ in phase with Θ_o for low values of ω but must lead Θ_o by an angle ϕ which increases as ω increases. The angle ϕ is 90° for $\omega = \omega_0$. The minimum magnitude of the Θ_i for the given curve occurs at a slightly lower value of ω. The corresponding plot of the variation of Θ_o for a given Θ_i for various values of ω is shown in Fig. 5.7.



(a) (b)

FIG. 5.8. Frequency response of quadratic component from the poles of the transfer function: (a) position of the poles; (b) vectors from poles to the $j\omega$ axis.

Another picture of the characteristics of a quadratic system is obtained by factoring the function in (5.13):

$$\frac{\Theta_o}{\Theta_i} = \frac{1}{(1 - j\omega/p_1)(1 - j\omega/p_2)} = \frac{-p_1}{j\omega - p_1}\,\frac{-p_2}{j\omega - p_2} \tag{5.14}$$

The poles p_1 and p_2 are the two values of s given in (3.51) and are, in general, complex numbers, as plotted in Fig. 5.8*a*. The designation ω_n for the j part of p_1 and p_2 is used to distinguish it from ω_0, which is the magnitude of p_1 or p_2, and ω, which is the general frequency variable.

The frequency response can be pictured by treating every factor in the final form of (5.14) as a vector, as shown in Fig. 5.8*b*. Thus $j\omega - p_1$ is a vector from the pole p_1 to a point on the j axis, $j\omega$. For $\omega = 0$ the vectors cancel; as ω increases, the vectors based on the poles p_1 and p_2 both rotate counterclockwise. The resonance of the system near $\omega = \omega_n$ is shown by the decreased length of the vector $j\omega - p_1$.

5.5. Open-loop Response

A control system involves several components acting in series so that the output of one is the input to the next. The total phase shift is thus the sum of the individual phase shifts, and the total attenuation is the product of the individual attenuations.

A typical control loop is shown in Fig. 4.1. The purpose of the control loop is to make the output shaft follow the input shaft. The error angle is measured as a voltage, which is amplified electronically and impressed across on the field of a generator. The generator voltage, which is proportional to the field current, is impressed across the motor terminals so that the motor drives the output to reduce the error. The analysis of the system is simplified by opening the loop, because then the output is not fed back to the amplifier. If a sinusoidal error voltage were impressed on the amplifier, all the other signals could be determined in vector form as shown in Fig. 5.9.

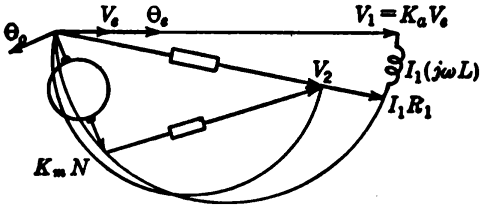


FIG. 5.9. Complete vector diagram of system.

The vectors are identified by symbols for the circuit elements themselves in order to emphasize the concept of the output of one component being the input to the next. The transfer function from field voltage to field current is a simple time delay, as discussed in Art. 5.3. The transfer function from motor voltage to resultant speed is the same form of function, but the time constant is much larger and so the phase shift is greater for a given frequency of signal. The transfer function from motor speed to motor angle is pure integration:

$$\frac{\Theta_o}{N}(j\omega) = \frac{\int N e^{j\omega t}\,dt}{N e^{j\omega t}} = \frac{(1/j\omega) N e^{j\omega t}}{N e^{j\omega t}} = \frac{1}{j\omega} \tag{5.15}$$

Thus for sinusoidal signals, integration produces 90° phase lag, regardless of frequency, and attenuation inversely proportional to frequency.

The system shown in Fig. 4.1 was shown in Art. 4.2 to be described by the block diagram in Fig. 4.3. The voltage v_f shown in Fig. 4.1 is a terminal value; the voltage v_1 is the open-circuit output voltage of the amplifier. The transfer function from θ_e to θ_o is the product of the individual transfer functions:

$$\frac{\Theta_o}{\Theta_e}(s) = \frac{V_e}{\Theta_e}(s)\,\frac{V_1}{V_e}(s)\,\frac{I_1}{V_1}(s)\,\frac{V_2}{I_1}(s)\,\frac{N}{V_2}(s)\,\frac{\Theta_o}{N}(s) \tag{5.16}$$

A symbol such as $(\Theta_o/\Theta_e)(s)$ is used for a transfer function from the signal θ_e to the signal θ_o because the signals involved are identified. A check that all the individual transfer functions are included is permitted by noting that the numerator of one function should be the denominator of the next function. The part of the symbol (s) indicates that the value of the function depends upon the value of s. For sinusoidal signals, $s = j\omega$, as noted in Arts. 4.3 and 5.3. Some functions such as

$$(V_e/\Theta_e)(s) = K_r$$

are simply constants, in this case the voltage obtained in the error-measuring system per unit error angle. Substituting in (5.16) the functions determined in (4.2), (4.3), and (5.15) gives

$$\frac{\Theta_o}{\Theta_e}(j\omega) = K_rK_a\frac{1/R_1}{1 + T_1j\omega}K_g\frac{1/K_m}{1 + T_2j\omega}\frac{1}{j\omega} \tag{5.17}$$

The product of all the constants in (5.17) is called the static loop gain K:

$$K = K_rK_a\frac{1}{R_1}K_g\frac{1}{K_m} \tag{5.18}$$

The gain K is the steady-state speed of the output shaft for a unit error angle θ_e. A dynamic loop gain may be defined as the ratio of amplitudes of motor angle and error angle for a specific value of ω. The dynamic gain is then the static gain times the attenuation of the frequency-dependent portions of the transfer function.

The transfer function $(\Theta_o/\Theta_e)(j\omega)$ is called the open-loop transfer function because the feedback loop is open for measurement of a quantity like static loop gain.

For a very low value of ω, the phase angle of the vector $(\Theta_o/\Theta_e)(j\omega)$ is 90°, owing to the integration of motor speed to angle, and the magnitude is very large. As the frequency is increased, the phase shifts due to T_1 and T_2 become significant, and the attenuation also increases. Finally at very high frequencies, the phase shift approaches 270° with complete attenuation.

5.6. Stability

> *A closed-loop system can maintain continuous oscillation at a frequency which makes the loop transfer function have a phase shift of* 180° *if the dynamic loop gain is unity. A system is said to be stable if any such oscillation will damp out. The loop frequency-response curve indicates stability if the* -1 *point is to the left of the curve as traced for increasing frequency. Nyquist's criterion is based on the position of the roots and will be proved in Chap.* 6.

The frequency-response method is based on the concept of steady-state conditions for sinusoidal signals. Thus, the characteristics of a system with the feedback loop open can be measured by impressing a sinusoidal signal on the amplifier input and measuring the voltage signal fed back.

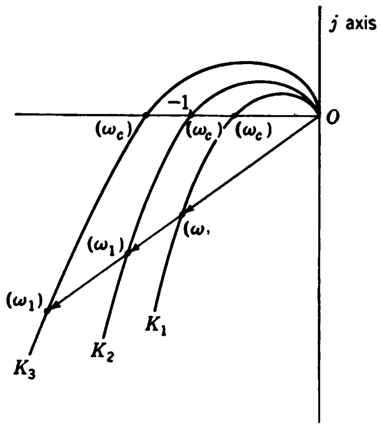


FIG. 5.10. Effect of loop gain on frequency response.

Any one curve in Fig. 5.10 is a vector plot of $(\Theta_o/\Theta_e)(j\omega)$ for a particular value of K and various values of ω. Increasing gain increases the length of a vector without changing the angle. The frequency ω_c is that which makes $(\Theta_o/\Theta_e)(j\omega)$ have an angle of 180°; the gain K_2 is that which makes the magnitude of $(\Theta_o/\Theta_e)(j\omega_c) = 1$. For the gain set at exactly K_2 an oscillation started at frequency ω_c will maintain itself at constant amplitude. In an actual test setup a value of gain slightly greater than K_2 will cause an oscillation to build up to an amplitude such that some component starts to saturate, thus reducing gain to K_2. For a system with a frequency-response curve of the simple shape shown in Fig. 5.10, the curve must cross the negative real axis inside the -1 point for the system to be stable.

The stability of a system in general *cannot* be judged by a *single point* along its frequency-response curve. Thus, a system with special circuits in the amplifier can have a loop transfer function as shown in Fig. 5.11 and be *stable*. The point at the frequency ω_2 has a phase shift of 180° and a dynamic loop gain greater than unity, just as ω_c has in Fig. 5.10 for K_3; yet the system is stable. Recall that each point on the curve is computed or measured for a *constant-amplitude* sinusoidal signal. If the curve misses the -1 point, the oscillation will either build up or decay. The transfer function ~~for a damped sine~~

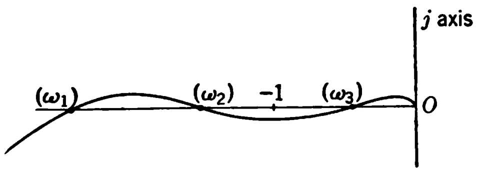


FIG. 5.11. A conditionally stable system.

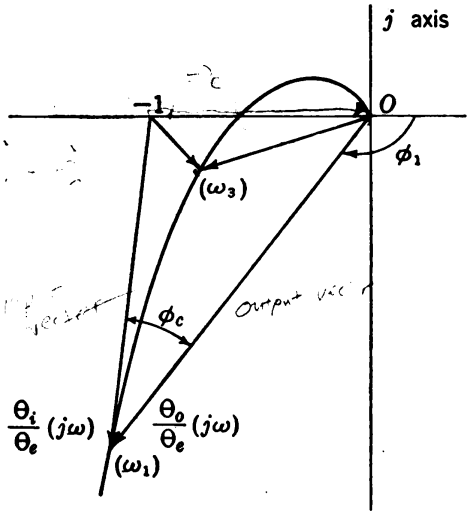
wave (Chap. 4) will, in general, have more phase shift and less attenuation than an undamped sine wave of the same frequency. Such a signal will make the loop transfer function equal to −1 if the −1 point is to the left of the curve as traced for increasing frequency. This concept is developed further in Chap. 6 to determine the roots of a system from its frequency-response characteristics.

The Nyquist criterion of stability is based on correlating the position of the roots on the complex plane with the frequency-response curve. A simple statement of the criterion is that a system is stable if the loop frequency-response curve does not "enclose the −1 point." In some cases, however, this is not clear unless the transfer function is plotted for the complete path of the variable s, which encircles the right half plane. The simple statement of the criterion applies only to systems in which the inner loops are stable. These considerations can be understood best perhaps by the proof itself, which is presented in Chap. 6.

5.7. Closed-loop Response

> *Loop gain is normally selected so that the system is stable by a sufficient margin that the resonant peak of the closed-loop frequency response is approximately* 1.4. *The closed-loop response can be obtained from the open-loop response by vector addition and the value of gain determined almost directly.*

A criterion of stability determines the limiting value of loop gain which is allowable. A margin of stability is necessary so that the closed-loop response does not have too high a resonant peak, as can be shown by a vector diagram (Fig. 5.12). The error vector is taken as unity, and the output vector at the frequency ω_1 is then the vector from the origin to the curve as shown. The input vector is the sum of the error and output vectors, which could be obtained by the usual parallelogram construction. A much simpler procedure, however, is to shift the tail of the $(\Theta_i/\Theta_e)(j\omega)$ vector to the −1 point with its head at the $(\Theta_o/\Theta_e)(j\omega)$ point on the curve. The ratio of output to input is then simply the ratio of vector lengths to the point on the curve. As higher values of ω are considered, the heads of the vectors follow along the curve



j axis
−1
O
ϕ_1
(ω_3)
ϕ_c
Output vec
$\frac{\Theta_i}{\Theta_e}(j\omega)$
$\frac{\Theta_o}{\Theta_e}(j\omega)$
(ω_1)

FIG. 5.12. Closed-loop frequency response as a ratio of vectors.

and pivot about the origin and the -1 point. The maximum ratio will be obtained at a frequency ω_m at which the $(\Theta_i/\Theta_e)(j\omega)$ vector is nearly perpendicular to the curve. Note that if gain were increased, the curve would approach the -1 point and the resonant peak would approach infinity.

The ratio of the output to the input at any frequency is called the magnification factor for the closed loop and is denoted by M. The determination of M is aided by drawing a set of curves in the complex plane at which the M values are constant. These curves prove to be circles with centers offset along the real axis. The procedure to obtain a particular value such as $M = 1.4$ involves some trial and error. The loop curve is first plotted to a convenient scale with a reasonable value of gain K_1. The position of the -1 point is then shifted until the ratio of the vector lengths $(\Theta_o/\Theta_e)(j\omega)/(\Theta_i/\Theta_e)(j\omega)$ is 1.4. The shift of the -1 point changes the scale of the drawing, and the gain K of the system is changed by the same factor. This construction is not explained in detail because the vector plot of the inverse function is preferable.

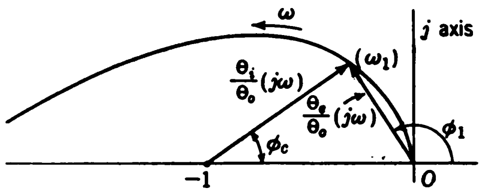


FIG. 5.13. Closed-loop transfer function using inverse transfer function.

The vector plot of the open-loop inverse transfer function $(\Theta_e/\Theta_o)(j\omega)$ differs from that of the direct transfer function in that the phase angles are reversed in sense and the magnitudes are the reciprocals. The system in Fig. 5.1, whose open-loop direct transfer function is shown in Fig. 5.12, has an inverse plot as shown in Fig. 5.13. The closed-loop function is now also in inverse form:

$$\frac{\Theta_i}{\Theta_o}(j\omega) = 1 + \frac{\Theta_e}{\Theta_o}(j\omega) \tag{5.19}$$

The vector $(\Theta_i/\Theta_o)(j\omega)$ is made larger by 1 than the $(\Theta_e/\Theta_o)(j\omega)$ vector by shifting its tail to the -1 point as shown. The reciprocal of the vector length $(\Theta_i/\Theta_o)(j\omega)$ is now the magnification factor M, and the phase shift from Θ_i to Θ_o is given by the angle ϕ_c.

The selection of gain such that the maximum value of M is 1.4 corresponds to making the minimum value of $(\Theta_i/\Theta_o)(j\omega)$ equal to 1/1.4, or 0.7. The function $(\Theta_e/\Theta_o)(j\omega)$ is first plotted using some reasonable value of gain K_1. Using the original scale of the plot, a circle arc is drawn about the -1 point with radius 0.7. If the arc misses the frequency-response curve, then the -1 point is shifted farther from the origin, thus changing the scale of the drawing, which corresponds to increasing the value of gain K. A circle arc about the new -1 point is drawn with a radius 0.7 based on the new scale. This process is repeated until the circle arc is just tangent to the frequency-response curve. For some shapes of the frequency-response curve no solution will be possible. An approximate solution for many systems, however, is that

$$\frac{\Theta_e}{\Theta_o}(j\omega) = 1\underline{/140^\circ}$$

for which $(\Theta_i/\Theta_o)(j\omega) = 0.7\underline{/70^\circ}$.

5.8. The Effect of a Feedback Function

The output signal may pass through a transfer function in the feedback path before being subtracted from the input. Stability is judged from the transfer function completely around the loop.

A system is often improved by mounting a tachometer on the output shaft to obtain a voltage proportional to the speed of the shaft. The signal to the amplifier is then determined by both the speed and the angle of the output shaft. The block diagram in Fig. 5.14 shows the net result after simplifications which eliminate voltages as separate signals. The differential equation represented in the feedback function is

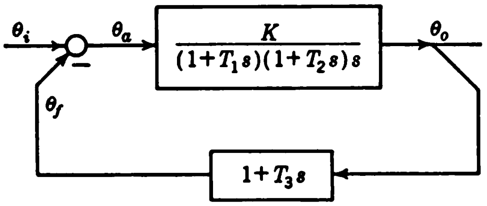


FIG. 5.14. A system with tachometer feedback.

$$\theta_f = \theta_o + T_3\frac{d\theta_o}{dt} \tag{5.20}$$

Stability is determined by the transfer function around the loop. The function $(\Theta_f/\Theta_o)(j\omega)(\Theta_o/\Theta_a)(j\omega)$ would have to be plotted in Fig. 5.10 for the stability criterion stated in Art. 5.6 to be applied. The signal to the amplifier is no longer the error between input and output and is, therefore, denoted as θ_a rather than θ_e. The relation between input and output is most easily expressed in terms of inverse functions.

$$\frac{\Theta_i}{\Theta_o}(j\omega) = \frac{\Theta_a}{\Theta_o}(j\omega) + \frac{\Theta_f}{\Theta_o}(j\omega) \tag{5.21}$$

For the particular case of the transfer functions given in the block diagram of Fig. 5.14, (5.21) becomes

$$\frac{\Theta_i}{\Theta_o}(j\omega) = \frac{(1 + T_1s)(1 + T_2s)s}{K} + (1 + T_3s) \tag{5.22}$$

If the gain K is high, then the first term is small, so that the inverse transfer function $(\Theta_i/\Theta_o)(j\omega)$ is approximately equal to the feedback function $1 + T_3s$. This result is true only for signals of a frequency less than that which makes the absolute value of the first term on the right in (5.22) small compared to the second term. This characteristic is occasionally used very effectively in improving the performance of a system by compensating for an undesirable transfer function of a component. A closed loop is used with the function in the feedback path of the loop approximately equal to that of the component. The overall transfer function of this loop is then approximately the reciprocal of the transfer function of the component.

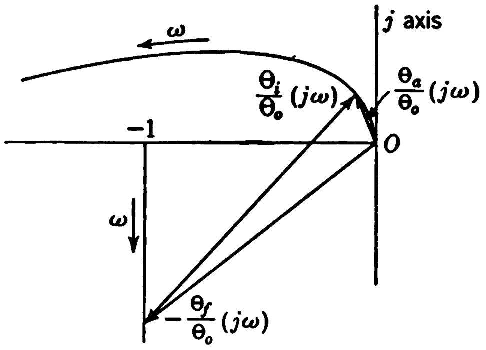

ω
j axis
$\frac{\Theta_i}{\Theta_o}(j\omega)$
$\frac{\Theta_a}{\Theta_o}(j\omega)$
-1
O
ω
$-\frac{\Theta_f}{\Theta_o}(j\omega)$

FIG. 5.15. Inverse system frequency response for system with tachometer feedback.

The actual transfer function of the loop in inverse form can be obtained by vector addition, as shown in Fig. 5.15 for the given system of Fig. 5.14. The approximation is valid to the extent that the vector $(\Theta_i/\Theta_o)(j\omega)$ is the same as the vector $(\Theta_f/\Theta_o)(j\omega)$.

5.9. Logarithmic Plots

Logarithmic plots permit simple straight-line approximations for a first-order transfer function. Multiplication of amplitude ratios can then be accomplished by addition of logarithms. The closed-loop response can be approximated readily, but reasonable accuracy requires the use of a special conversion chart.

The transfer function of a first-order system can be described by separate plots of amplitude and phase versus frequency. Use of a logarithmic scale for frequency permits simple straight-line approximations for each curve, as shown in Fig. 5.16. The abscissa is plotted to a logarithmic scale with the numerical value of ω given. The logarithm of the magnitude of the transfer function is commonly expressed in *decibels*, a decibel being defined as twenty times the common logarithm of the magnitude. The following approximations are used:

$$1 + j\omega T = 1 \qquad \text{for } \omega T < 1 \tag{5.23}$$

$$= j\omega T \qquad \text{for } \omega T > 1 \tag{5.24}$$

These approximations produce straight lines on the logarithmic plot, as shown in Fig. 5.16*b*. The actual value of the magnitude is $\sqrt{2}$ at $\omega T = 1$, for which the logarithmic value is

$$\text{db} = 20 \log_{10} \sqrt{2} = 20 \times \tfrac{1}{2} \times 0.301 \cong 3 \tag{5.25}$$

The approximation is surprisingly accurate at values of ω removed from the break point at $\omega T = 1$. For example, at $\omega T = 2$, the true magnitude is 7 db compared to an approximate value of 6 db, an error of only 1 db, or 12 per cent.

The straight-line approximation for phase shift shown in Fig. 5.16*a* is surprisingly accurate also. The approximation is a straight line from 0° to 90°, starting at $\omega T = 0.1$ and ending at $\omega T = 10$. The true phase-shift curve crosses this straight-line approximation three times and is never in error by more than 6°. The approximation is particularly convenient because the frequency range of 10 to 1, called a decade, is easy to identify in plotting.

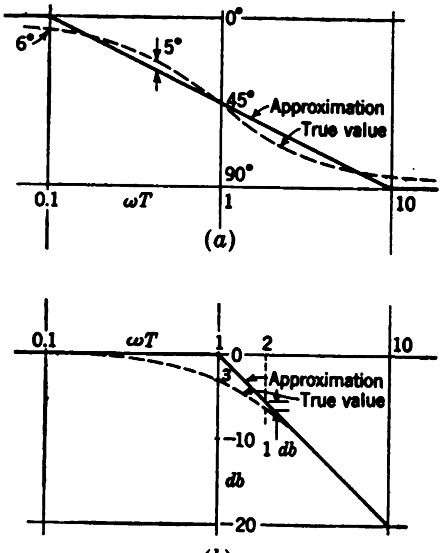


FIG. 5.16. Frequency-response approximations for transfer function $1/(1 + Ts)$: (*a*) phase angle; (*b*) attenuation.

Note that the presence of the term $1 + j\omega T$ in the denominator of the transfer function causes the phase shift to be negative and the decibel value of the magnitude to be negative. The scale in Fig. 5.16 is such that the magnitude approximation has a slope of -1 and the phase-angle approximation has a slope of $-\frac{1}{2}$.

Consider now the construction of a logarithmic plot for an open-loop transfer function of the form

$$\frac{\Theta_o}{\Theta_e}(j\omega) = \frac{K}{(1 + T_1 j\omega)(1 + T_2 j\omega)j\omega} \tag{5.26}$$

Figure 5.17 is based on $K = 1$, $1/T_1 = 1$, and $1/T_2 = 4$.

The straight-line approximation to the magnitude curve will first be established. For $\omega < 1/T_1$, the function is simply $1/\omega$, which is plotted as a line with slope of -1 having an ordinate of 0 db at $\omega = 1$. For $1/T_1 < \omega < 1/T_2$, the function is approximated by $1/\omega^2 T_1$, which is plotted as a line with a slope of -2. For $\omega > 1/T_2$, the function is approximated by $1/\omega^3 T_1 T_2$, which is plotted as a line having a slope of -3. A better approximation to the true magnitude can now be obtained by sketching a curve which is asymptotic to the straight lines but falls below the *break* points $1/T_1$ and $1/T_2$ by 3 db.

The phase-angle curve is likewise first approximated by straight-line segments. For $\omega < 0.1/T_1$, the phase shift is 90°. At $\omega = 0.1/T_1$, a decade ahead of the break frequency $1/T_1$, the angle function takes on a slope of $-\frac{1}{2}$, in accordance with the approximation shown in Fig. 5.16*b*.

At $\omega = 0.1/T_2$, a decade ahead of the second break frequency, the slope of the line doubles, becoming -1. This slope is continued to a decade past the first break frequency, $\omega = 10/T_1$, after which the slope becomes $-\frac{1}{2}$. The final asymptote is 270° after $\omega = 10/T_2$.

The stability of the system with the feedback loop closed is judged by Nyquist's criterion, although some imagination is required at first. The function has a magnitude of 1 when its decibel value is zero. At this frequency, the phase shift is less than 180°. A vector plot of this simple

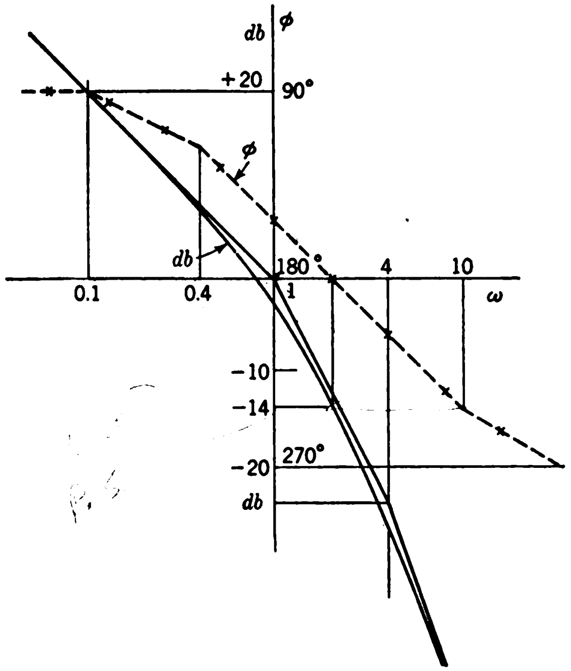


FIG. 5.17. Frequency-response plot for $K/(1 + T_1s)(1 + T_2s)s$, in which $1/T_1 = 1$, $1/T_2 = 4$, $K = 1$.

function would thus come inside the -1 point, and hence the system would be stable. A gain factor $K = 2$ can be introduced into the transfer function in (5.26) simply by lowering the 0-decibel line of Fig. 5.17 by 6 db. The 0-decibel line can be lowered 14 db before *dynamic loop gain* is unity at the frequency which produces a phase shift of 180°. The system would then be unstable, making $K = 5$ the limit.

The gain K is usually chosen so that the maximum magnification of a signal at any frequency from input to output of the closed-loop system is approximately 1.4. A vector plot such as Fig. 5.12 will show that this maximum magnification often occurs when the open-loop function is nearly unity in magnitude at an angle of 140°. On a logarithmic plot the frequency ω_1 at which the open-loop phase angle is 140° is located as shown in Fig. 5.18. The 0-db line is then adjusted to intersect the magnitude curve at this frequency. The closed-loop transfer function then has a phase angle of 70° and a magnitude of 1.4, corresponding to a decibel value of 3.

Denoting the open-loop function $(\Theta_o/\Theta_e)(j\omega)$ as $KG(j\omega)$, the closed-loop function can be shown to be

$$\frac{\Theta_o}{\Theta_i}(j\omega) = \frac{KG(j\omega)}{1 + KG(j\omega)} \tag{5.27}$$

For $\omega \ll \omega_1$, $KG(j\omega) \gg 1$, and therefore $(\Theta_o/\Theta_i)(j\omega) \cong 1$. For $\omega \gg \omega_1$, $KG(j\omega) \ll 1$, and therefore $(\Theta_o/\Theta_i)(j\omega) \cong KG(j\omega)$. The remainder of the closed-loop transfer-function curves are sketched on the basis of the above limits. The middle portion of the curve is usually needed to a greater accuracy than that permitted by just one point; this greater accuracy can be achieved by use of a special conversion chart.

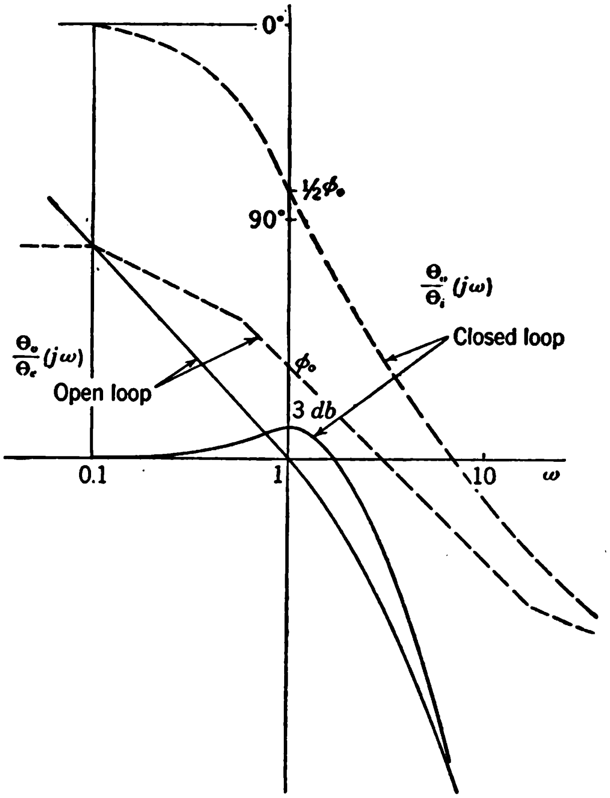


Fig. 5.18. Approximate closed-loop frequency response for the loop function given in Fig. 5.17.

The conversion from open-loop to closed-loop frequency response depends upon the length and phase angle of the open-loop-response vector, not upon the frequency at which this occurs. The length of the vector is identified in Fig. 5.17 logarithmically in decibel value; thus, it is convenient to make the conversion to closed-loop response in terms of these units. The open-loop characteristic can be plotted as a contour of decibels versus phase angle, as shown in Fig. 5.19.

The plotting of the contour curve from the data in Fig. 5.17 is facilitated by making the scale of the two figures the same. Plotting the contour on a transparent sheet makes it possible to mark the lengths directly from Fig. 5.17 without having to read values and replot points. This transparent sheet, or *overlay*, has a 45° line for convenience in shifting from vertical lengths to horizontal lengths. Thus, the overlay is shifted horizontally until the 45° line intersects the phase-angle curve (point 1). The abscissa on the overlay is then the same length as the ordinate for the phase angle. A vertical line through point 1 intersects the magnitude curve at point 2; this point is marked on the overlay, thus giving decibels versus phase angle on the overlay. A similar procedure determines other points along the curve as indicated. This overlay now describes the open-loop frequency response of the system in the range for which $KG(j\omega)$ is of the same order of magnitude as unity.

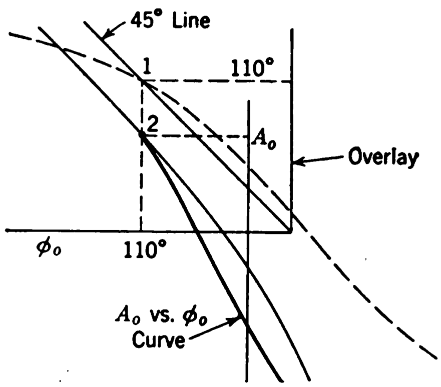
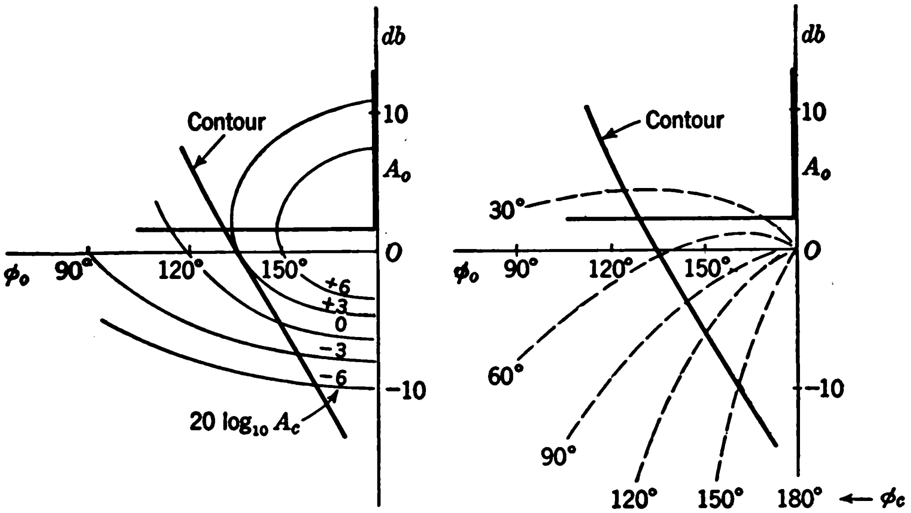


FIG. 5.19. Plot of attenuation versus phase angle obtained from Fig. 5.17 by using an overlay.



(*a*) (*b*)

FIG. 5.20. Conversion charts from open-loop frequency response to closed-loop frequency response: (*a*) attenuation; (*b*) phase angle.

The overlay can now be placed on a special conversion chart to determine the closed-loop response of the system, as shown in Fig. 5.20. The overlay can be shifted vertically to increase the value of the gain K from the original value of 1. A maximum magnification of 3 db can be achieved by shifting the overlay until the contour becomes tangent to the 3-db curve of the conversion chart in Fig. 5.20*a*. The allowable gain increase is 2 db, which is nearly the same as the value obtained by the approximate method in Fig. 5.18. Other points for closed-loop response can be read from these conversion charts to obtain a reasonably accurate curve in the region of the resonant peak. Other methods of finding closed-loop response from open-loop response can, of course, be used, including the plotting of a vector curve from the logarithmic data.

5.10. Use of Integral Circuit

An integral circuit permits static gain to be increased by providing a corresponding attenuation of the signal. The time constants of the circuit are selected so that little phase lag is introduced at the frequency at which the system tends to oscillate.

An integral circuit and its characteristics in logarithmic form are shown in Fig. 5.21. The transfer function of the circuit can be set up as the ratio of impedances:

$$\frac{V_o}{V_i}(s) = \frac{R_2 + 1/C_2s}{R_1 + R_2 + 1/C_2s} = \frac{1 + T_2s}{1 + T_1s} \tag{5.28}$$

The ratio of time constants is commonly taken to be 10:1, so that the attenuation provided by the circuit for frequencies above $10/T_2$ is 20 db. A phase lag is necessarily present in the frequency range near the break frequencies, $1/T_1$ and $1/T_2$. This lag need not complicate the stability problem, however, if $10/T_2$ is made to occur at or below the frequency for which the magnitude of the open-loop function is nearly 1, as illustrated in Fig. 5.22.

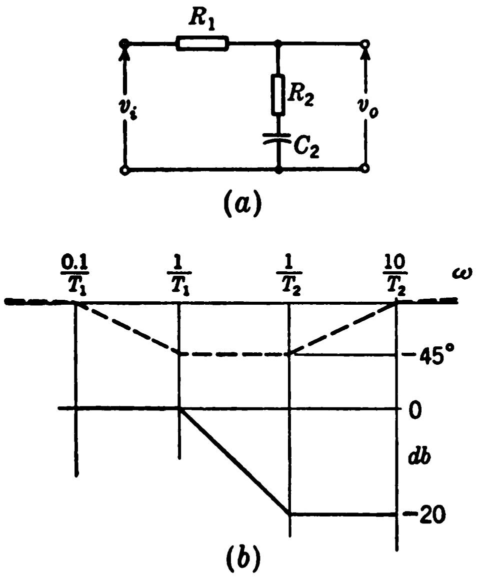


FIG. 5.21. An integral network: (*a*) the circuit; (*b*) the frequency-response characteristics.

The light lines, solid and dashed, on this figure correspond exactly to those of Fig. 5.17. The heavy lines show the characteristic of the open-loop transfer function when the integral circuit is in series and the amplifier gain is raised by 20 db. The phase shift of the new system, shown by the heavy dashed line, is now 135° over a wide range of frequency, but the phase shift at the frequency of 0-db crossover of the magnitude curve is not increased by more than 6°. The magnitude curve starts 20 db higher at the low-frequency range, but the attenuation provided by the integral circuit brings the curve back to its former characteristic for frequencies above $1/T_2$.

The closed-loop response is not appreciably affected by the addition of the integral network. At low frequencies, the response is approximately 1, as before. The one point established at 0-db crossover will be nearly the same as before because the phase shift is nearly the same. Finally, for high frequencies, the characteristics of the system are the same.

An integral circuit is ideally used in applications where a low error is required during time intervals in which the input is stationary or has a constant velocity. The voltage needed across the motor, and hence v_o in Fig. 5.21, is then essentially constant. The capacitor thus has time to achieve nearly full charge, which eliminates the attenuation of the circuit, permitting a lower error voltage.

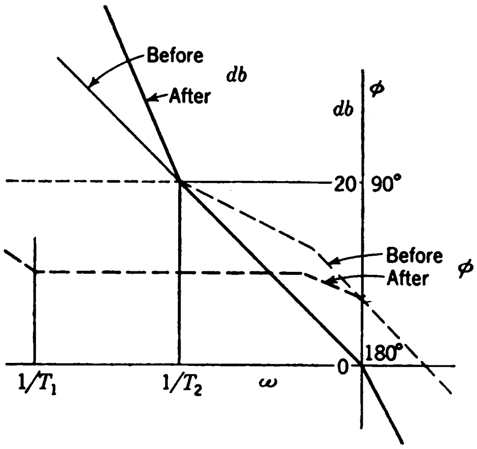


FIG. 5.22. Modification of the system in Fig. 5.17 by addition of the integral network in Fig. 5.21.

5.11. Use of Lead Circuit

A lead circuit produces a phase lead over a particular frequency band, with a corresponding amplification of the signal necessarily present. The time constants of the circuit are selected so that the maximum lead is attained near the frequency for which the remainder of the open-loop transfer function is 180°. *This circuit permits the loop gain to be increased, and this in turn raises the frequency of the resonant point of the closed-loop response.*

A lead circuit and its characteristics in logarithmic form are shown in Fig. 5.23. The transfer function from v_1 to v_2 can be written as a ratio of impedances:

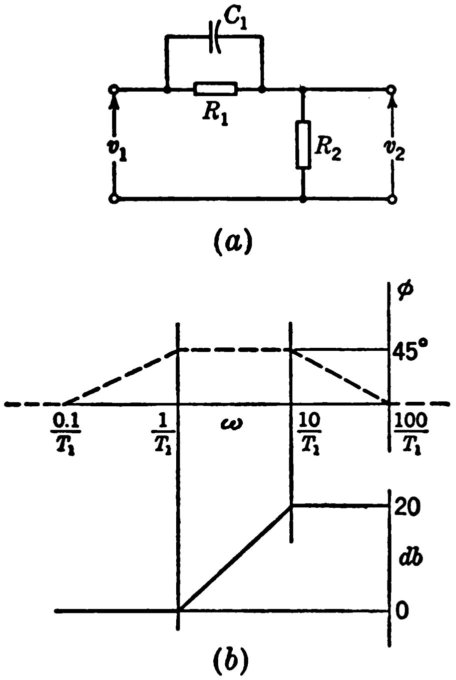
$$\frac{V_2}{V_1}(s) = \frac{\alpha(1 + T_1 s)}{1 + \alpha T_1 s} \tag{5.29}$$

in which

$$T_1 = R_1 C_1 \qquad \alpha = \frac{R_2}{R_1 + R_2}$$

The form of this transfer function is seen to be the same as that for the integral circuit, but the numerator time constant is now larger than that of the denominator. The logarithmic characteristics are thus similar but are inverted. The change in magnitude with frequency, here denoted as α, is again taken to be 10 for convenience.

The effect of this circuit on the system of Fig. 5.17 is shown in Fig. 5.24. The characteristics of the system before and after the addition of the lead circuit are identified. The lead break frequency $1/T_1$ is made equal to the frequency at which the original system had a magnitude of 0 db. Note that the phase lag is decreased by the trapezoidal pattern introduced by the lead circuit. A frequency of ω_a is now required to make the total phase shift 140°, as seen at point P. This result means that the gain can be increased by 10 db to make the 0-db crossover of the magnitude curve occur at this frequency (point P). The closed-loop response of the system will have essentially the same shape as that of Fig. 5.18, except that the resonant frequency will be increased.



(a)

(b)

FIG. 5.23. A lead network: (a) the circuit; (b) the frequency-response characteristics.

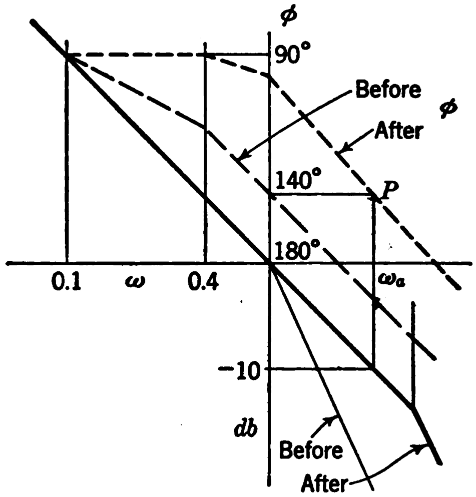


FIG. 5.24. Modification of system in Fig. 5.17 by addition of the lead network in Fig. 5.23.

A lead circuit is applicable in a closed-loop system in which the cutoff frequency must be raised. The cutoff frequency is the frequency above which the system attenuates the signal. It requires an increase in amplifier gain to compensate for the attenuation of low-frequency signals in the lead circuit. The use of this circuit by itself has the disadvantage of amplifying high-frequency signals, which are often spurious signals, or *noise*, which should be attenuated.

5.12. The Question of Transient Response

The frequency-response method avoids the problem of finding the roots of the characteristic equation by opening the feedback loop and describing the system by its steady-state response to sinusoidal signals. Stability of the system is determined, methods of improving the system are suggested, and the final result is the frequency response of the closed-loop system. If transient response is needed, conversion of data or a complete new solution is necessary.

The application of frequency-response concepts to control systems arises from the similarity of the closed-loop system to a feedback amplifier. The feedback amplifier is involved in systems in which the signals are usually considered to be sinusoidal, and the frequency response of the amplifier is, therefore, exactly the information desired. A control system, on the other hand, is subject to a wide variety of inputs such as are involved in tracking a moving target. The transient response is then needed to determine some quantities, *e.g.*, the time required to lock in on a new target. This problem can be solved in any one of several ways; the preferred approach will depend to some extent upon the individual experience of the worker.

One approach is to use the frequency-response method to determine the optimum system, using criteria such as maximum magnification of 3 db, and then to set up and solve the differential equation of the system to find the transient response. Another approach is to use the frequency response of the closed-loop system to obtain the transient response by some form of evaluation of a Fourier integral. Another approach is to correlate the frequency response and transient response of many systems, using test results or machine calculation.

All these methods entail concepts different from those involved in the frequency-response method itself, and thus this problem is considered in the next chapter.

CHAPTER 6

TRANSIENT RESPONSE FROM FREQUENCY RESPONSE

The frequency-response method is valuable because it suggests the changes necessary in a system to improve its performance. Similarly, a method of finding the transient response from frequency-response data is most valuable in so far as it suggests the changes needed in the frequency-response curve to produce a desired transient response. On this basis the methods studied in this chapter are those which determine the zeros and poles of the system transfer function. The poles of this function are the roots of the characteristic equation of the system and describe the transients of the form e^{st} which can exist in the system. The amplitudes of terms can be determined easily using the results of the Laplace transform in Chap. 4.

Two methods of finding the system transfer function in factored form are considered: curve fitting and conformal mapping. Each method involves trial and error, but each one is reasonably fast. Curve fitting is particularly appropriate for the logarithmic plots versus frequency. Conformal mapping is best suited to the vector plot of the inverse loop transfer function. It will be shown that a pattern of curvilinear squares can be sketched using the vector plot as a boundary to determine the complex numbers which make the loop transfer function equal to -1. Study of this method is *not* essential to understanding the root-locus method in Chap. 7.

6.1. Asymptotes on Logarithmic Plots

> *A system whose logarithmic response curve can be approximated by straight-line asymptotes has a transfer function which is a product of terms of the form* $1 + Ts$. *The phase-shift characteristic provides a check on the function.*

Finding the transfer function of a system with known frequency-response characteristics is the reverse problem from that considered in Chap. 5. Therefore, it is appropriate to illustrate the procedure using the logarithmic plot developed in Fig. 5.17.

The curves are known in this case by direct calculation, but the same type of curves might be obtained by experimental measurement of equipment. Finding the asymptotes is a trial-and-error process; the first trial is usually best made on the basis of the attenuation characteristic alone, as shown in Fig. 6.1.

The low-frequency asymptote with a slope of unity, or 20 db/decade, is easy to establish accurately. The location of the break points for the succeeding asymptotes is not so certain. The frequency values at the break points are the reciprocals of the time constants in the transfer function. The approximation to the attenuation characteristic thus determines the transfer function, which in turn determines a phase-shift characteristic. For this first trial, the actual phase shift of the system is less than that of the assumed function. This deviation can be corrected by shifting the break points to slightly higher frequencies. Successive trials are made until the attenuation and phase-shift curves are both matched. The location of the break points may not be critical; if the errors in mismatch of the frequency characteristics are small, then the error in the transient-response plot will also be small.

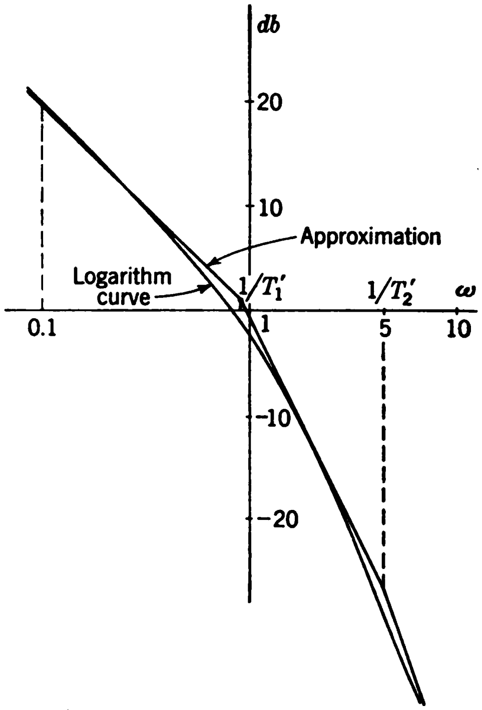


FIG. 6.1. Asymptotic approximation to a frequency-response curve.

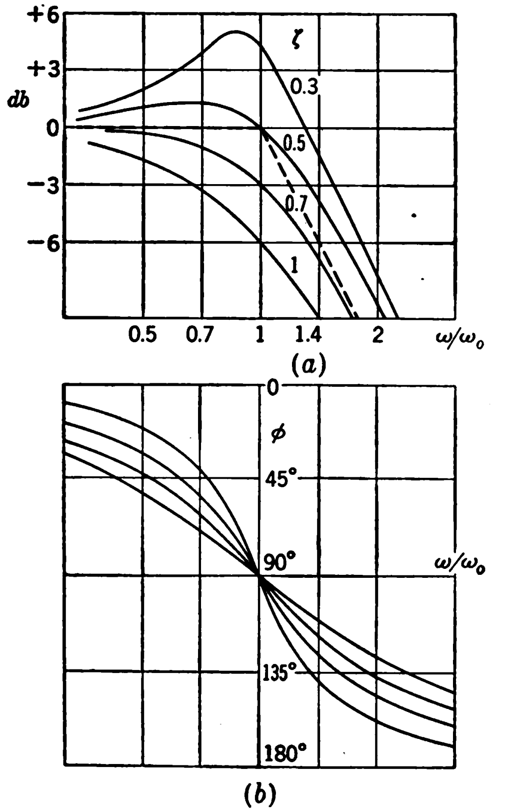
Frequency-response data obtained experimentally present two problems: errors in measurement and unexpected factors in the transfer function. Phase-shift measurements, for example, are usually less accurate than attenuation measurements; therefore, more attention should be devoted to matching the attenuation curve than is given the phase-shift curve. On the other hand, the phase-shift angle at the high-frequency end of the data is often greater than that expected from the attenuation characteristic by more than can be attributed to experimental error. Recall, however, that a transfer function of the form $1/(1 + Tj\omega)$ introduces 15° phase shift for $\omega T = \frac{1}{4}$, whereas the magnitude of the function is just 1.03. In comparatively rare cases, the phase shift in the middle of the range of measurements may be greater than appears to be consistent with the attenuation curve. Such a characteristic can arise with bridge circuits, having a transfer function such as

$(1 - Ts)/(1 + Ts)$. The numerator term in this function produces amplification and phase lag at frequencies above $1/T$. Frequency-response characteristics having sharper changes in attenuation and phase shift than can be matched by straight-line asymptotes indicate the presence of quadratic terms in the transfer function of the system.

6.2. Quadratic Functions

> *A closed-loop function usually has a resonant peak in its attenuation characteristic at some frequency. Curve fitting is aided by use of a set of curves for quadratic functions with the damping ratio as parameter.*

Consider the closed-loop frequency response determined in Fig. 5.18. The resonant peak in the amplitude curve cannot be matched by any combination of straight-line asymptotes. The transfer function has a quadratic factor for which the poles are complex numbers. The characteristics of quadratic functions can be depicted by a set of curves like those shown in Fig. 6.2.



(a)

(b)

Fig. 6.2. Characteristics of a quadratic function $1/(1 + 2\zeta s/\omega_0 + s^2/\omega_0^2)$: (a) attenuation; (b) phase angle.

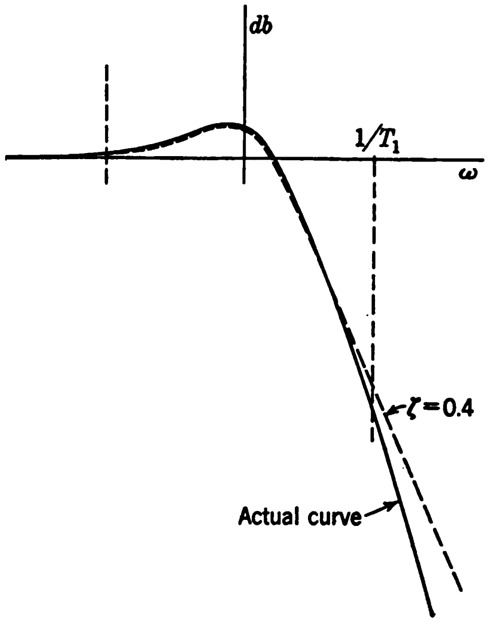


Fig. 6.3. Matching a quadratic curve to a frequency-response curve.

These curves are in completely dimensionless form: ω/ω_0 is the ratio of frequency to the undamped natural frequency of the quadratic, ω_0; ζ is the damping ratio, or the ratio of damping constant σ to ω_0. A quadratic system with a given damping ratio has a characteristic curve which has the same shape regardless of ω_0. The actual response curve can thus be placed on the quadratic curves at any position along the ω axis for best fit, as shown in Fig. 6.3. The given curve appears to have the best fit with the quadratic curve with a damping ratio of 0.4. The difference between the two curves can then be plotted to obtain the characteristic of the remaining factors of the system transfer function. In this case a first-order factor with break point at $1/T_1$ appears to match the remaining characteristic. The transfer function is thus completely determined, and the transient response of the system can be written by inspection using the inverse Laplace transform or the method described in Chap. 4.

If the effect of changing gain is to be studied, the closed-loop function must be plotted from the loop function and the curve-fitting procedure repeated. A method of determining the poles of the closed-loop function directly from the loop function is desired. The Nyquist criterion of stability ensures, on the basis of the open-loop frequency response, that these poles have negative real parts. A proof of this criterion based on vectors in the complex s plane will suggest a conformal-mapping means by which the poles themselves can be determined.

6.3. Nyquist Criterion of Stability

> *The Nyquist criterion, for simple systems, states that the system is stable if the curve of the open-loop transfer function versus frequency does not enclose the* -1 *point. The precautions which are necessary for applying the criterion to complicated systems can best be appreciated by understanding the proof of the criterion itself.*

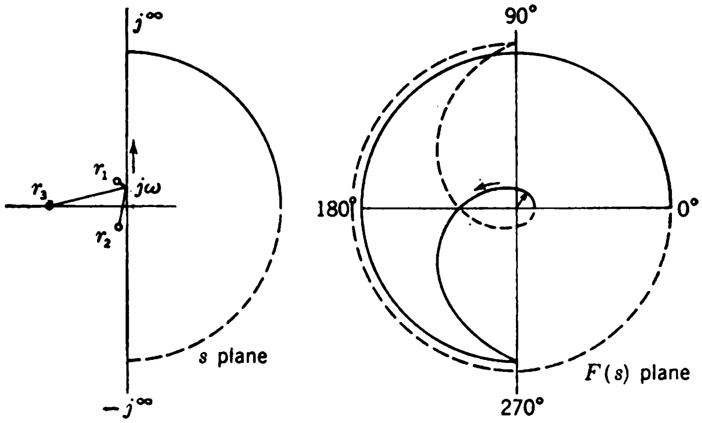
The proof of Nyquist's criterion is very simple in principle, as will be demonstrated in this section. Consider the simple system shown in Fig. 5.1. The roots are the values of s which satisfy the characteristic equation:

$$\frac{\Theta_i}{\Theta_o}(s) = 1 + \frac{\Theta_e}{\Theta_o}(s) = 1 + \frac{(1 + T_1s)(1 + T_2s)s}{K} = 0 \tag{6.1}$$

Assume that these roots are known, and designate them as r_1, r_2, and r_3. The characteristic equation can then be rewritten as

$$\begin{aligned} F(s) = \frac{\Theta_i}{\Theta_o}(s) &= \left(1 - \frac{s}{r_1}\right)\left(1 - \frac{s}{r_2}\right)\left(1 - \frac{s}{r_3}\right) \\ &= \left(\frac{-r_1 + s}{-r_1}\right)\left(\frac{-r_2 + s}{-r_2}\right)\left(\frac{-r_3 + s}{-r_3}\right) \end{aligned} \tag{6.2}$$

This function $F(s)$ can be sketched versus frequency as a parameter very easily by using the vector relations shown in Fig. 6.4. Note that the roots r_1 and r_2 are assumed to be a complex conjugate pair and the third root is along the negative real axis. For a particular value of $j\omega$, s is a vector with its head at the $j\omega$ point and the tail at the origin. The factor $-r_1$ is a vector from the root r_1 to the origin. The vector $j\omega - r_1$ is the sum of these two vectors, or a vector from the r_1 point to the $j\omega$ point. Similarly other factors in the numerator of (6.2) are vectors from the other roots to the $j\omega$ point. The product of these vectors is a vector whose phase angle is the sum of the phase angles and whose magnitude is the product of the magnitudes. The denominator factors are constant and have no net phase angle. For a particular value of s as shown in Fig. 6.4a the function in (6.2) is the vector shown in Fig. 6.4b.



(a) (b)

Fig. 6.4. Mapping the boundary of the right half plane of the s plane in the $F(s)$ plane: (a) the vectors from the poles of $F(s)$; (b) the vector plot of $F(s)$.

Note that, if the s point starts at the origin and moves up the axis, all the vectors with heads at the s point will rotate while pivoting at the roots. The function $F(s)$ will be the continuous curve shown in Fig. 6.4b. Note that, as the s point approaches $j\infty$, each variable vector approaches $j\infty$, making the function approach 270°. This part of the curve determines the stability of the system, although the full proof requires that the path of the s point be completed around the right half plane. When the s vector moves in an arc back to the real axis, all the vectors in the function swing from 90° to 0°, making the function curve swing back through 270° as shown. As the path of the s point continues around the lower right half quadrant, each of the function vectors is the same as before in every respect, except that the phase angles are opposite in sign. The function curve thus becomes the mirror image about the horizontal axis of the curve for s tracing around the upper right half quadrant.

The vector from the origin to the function curve in Fig. 6.4b has no net rotation for s tracing a path which encloses the right half plane. If the roots r_1 and r_2 are in the right half plane, however, the vectors $s - r_1$ and $s - r_2$ would each make a complete revolution and the $F(s)$ function vector would make two revolutions.

The $F(s)$ function curve, however, can be obtained from the loop transfer function by adding 1 as seen in Eq. (6.1). If $(\Theta_e/\Theta_o)(s)$ is plotted about 0 as the origin, then the curve for $(\Theta_i/\Theta_o)(s)$, or $F(s)$, is obtained simply by shifting the origin of the curve to the -1 point. The number of roots lying in the right half plane then corresponds to the number of rotations of the vector from the -1 point to the inverse loop frequency-response curve. The critical part of the criterion is simply that the curve for the inverse function must swing outside of the -1 point.

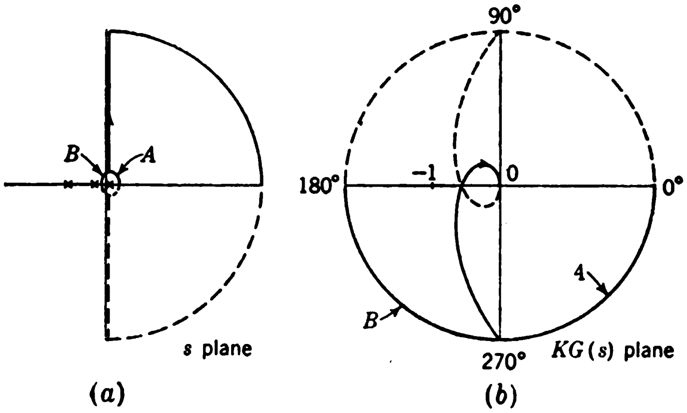


FIG. 6.5. Effect of indentation around the origin in the s plane on the plot in the $KG(s)$ plane.

The Nyquist criterion using the direct loop transfer function is more common and illustrates the care which must be exercised in using the criterion. The plot of the direct transfer function as viewed from the -1 point is

$$\frac{\Theta_i}{\Theta_e}(s) = 1 + \frac{\Theta_o}{\Theta_e}(s) = \frac{K + (1 + T_1s)(1 + T_2s)s}{(1 + T_1s)(1 + T_2s)s} = 1 + KG(s) \qquad (6.3)$$

Actually only the numerator is of interest in locating the roots; the presence of the denominator factors is just a necessary evil. In particular, these latter factors may cause rotation of the function vector in (6.3), which must be taken into account.

Consider the plot of the direct transfer function shown in Fig. 6.5. For s moving up the $j\omega$ axis from near the origin to $j\infty$, the direct function curve moves from $-j\infty$ to approach the origin along the $+j$ axis. The function is zero for the entire sweep of the s point around the right half plane. The motion of the s point up the $-j\omega$ axis produces a function curve which is the mirror image of the previous curve. At this stage of plotting, one cannot say whether or not the vector from the -1 point to the function curve makes a complete revolution. It all depends upon whether the curve is closed at infinite radius to the right or to the left; the path of the function curve in turn depends upon the path chosen for the s point near the origin. For path A, to the right of the origin, the function curve closes around the right half plane. For path B, to the left of the origin, the function curve closes around the left half plane. In this latter case the vector to the function curve rotates counterclockwise one revolution, corresponding to the s vector in the denominator of (6.3) making one revolution clockwise. The number of rotations of a function vector is equal to the difference between the number of zeros and the number of poles of the function which lie in the right half plane. When using the Nyquist criterion for a multiple-loop system, therefore, the function being plotted should be expanded as in (6.3) to determine if any poles are in the right half plane, for they cancel rotations due to zeros in that half plane.

6.4. Finding Roots by Conformal Mapping

A grid of squares in the s plane maps into a grid of curvilinear squares in the function plane. Conversely, a grid of curvilinear squares can be sketched in the function plane, using the frequency-response curve as a boundary, to determine the complex values of s which make the loop transfer function equal to -1.

The vector plots used in proving Nyquist's criterion provide a rapid method of estimating the value of a function of s for complex values of s as well as for values along the $j\omega$ axis. Thus consider the problem of plotting the value of the inverse system function in (6.2) for values of s in the region of r_1. In particular, a grid of squares in the s plane will be shown to map into a grid of curvilinear squares in the function plane, as shown in Fig. 6.6.

Note first that for $s = r_1$ in the s plane, the function in (6.2) must be zero because the factor $s - r_1$ is zero. In the region near $s = r_1$ the other factors in (6.2) are nearly constant, which means that the changes in the function are due almost entirely to the changes in the vector $s - r_1$. A grid of squares in the s plane would thus be expected to map into a grid of squares in the function plane which are only slightly distorted. The plot in the function plane (Fig. 6.6*b*) is somewhat similar to plots of static electric or magnetic fields. Such plots are characterized by lines which cross at right angles and by curvilinear squares which, if sufficiently subdivided, become true squares. These properties can be shown to be true for the function plot by considering small increments, as shown in Fig. 6.7.

The increment in the function $F(s)$ in (6.2) can be expressed as

$$\Delta F(s) = \frac{dF(s)}{ds}\,\Delta s \tag{6.4}$$

The derivative $dF(s)/ds$ has a perfectly definite value for the given value of s. The increment in the function therefore depends entirely upon the increment in s. Consider the particular point $j\omega_1$. Usually the increment in s is $\Delta j\omega$, for which the function increases by an amount along the frequency-response curve as shown. But if the increment in the variable is $-\Delta\sigma$, perpendicular to $\Delta j\omega$, then the increment in the function must be perpendicular to the frequency-response curve. Similarly, if the magnitudes of the increments in s are taken to be the same, then the magnitude of the increments in the function must be the same. Thus if a pattern of small squares is made in the s plane, this grid will *map* into a curvilinear grid in the function plane with each element being essentially a square. Such a map is said to be conformal. A function of s can be conformally mapped in any region for which the derivative of the function has a definite value.

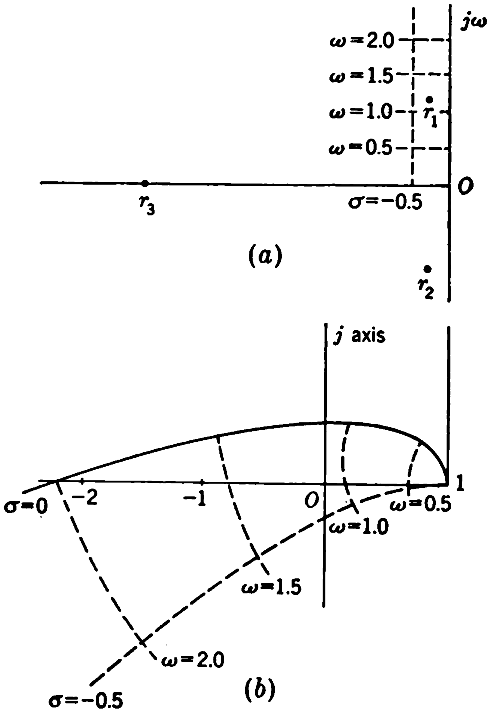
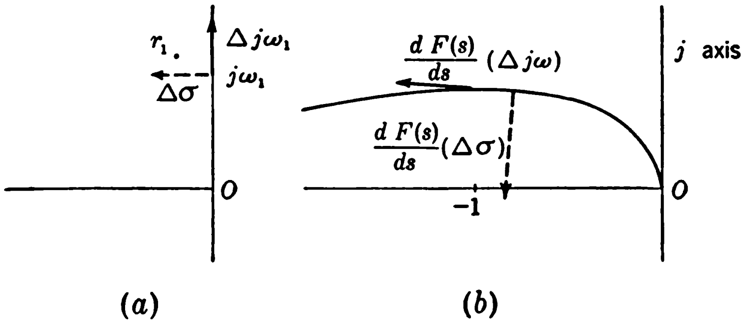


FIG. 6.6. Conformal map: (*a*) square grid in the s plane; (*b*) the $F(s)$ plane.


r_1
$\Delta\sigma$
$\Delta j\omega_1$
$j\omega_1$
O
$\frac{dF(s)}{ds}(\Delta j\omega)$
$\frac{dF(s)}{ds}(\Delta\sigma)$
−1
O
j axis

(*a*) (*b*)

FIG. 6.7. Proof of conformal mapping: (*a*) increments in the s plane; (*b*) corresponding increments in the $F(s)$ plane.

The conformal-mapping property of a transfer function provides a means by which the values of the complex roots may be estimated from the frequency-response curve. The vector plot for the inverse loop function in Fig. 5.1 is shown in Fig. 6.8 for $K = 1$. The roots of the characteristic equation are the value of s which make this function equal to -1. The value of the function for complex values of s can be estimated by sketching a conformal map using the frequency-response curve as a boundary. Interpolation gives the value of the root as

$$r_1 = -0.4 + j0.85$$

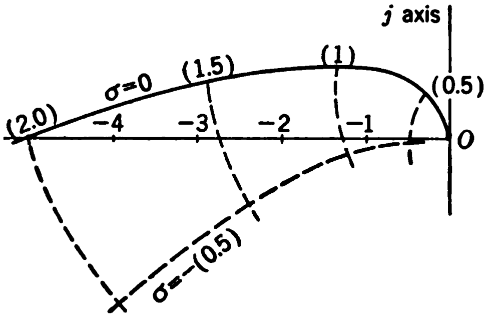


Fig. 6.8. A conformal map sketched from a frequency-response curve.

6.5. Amplitude of Transient Terms

The amplitude of a transient term can be expressed in terms of a derivative of the transfer function. This derivative can be evaluated directly on the conformal map at the -1 *point.*

The conformal-mapping process determines one pair of complex roots of the characteristic equation of the system. Previous experience with the frequency-response curves suggests that this pair is the dominant pair of roots of the system. If a step input were applied, for example, the main portion of the output transient would be a damped oscillation with frequency and damping rate as specified by the known roots. This conjecture could be verified if the amplitude of the transient term could be determined.

The classic procedure would require that all the roots be determined, and the solution set up in general form:

$$\theta_o = A_1 e^{r_1 t} + A_2 e^{r_2 t} + A_3 e^{r_3 t} + \cdots \tag{6.5}$$

The initial conditions are then used to obtain a set of simultaneous equations, the solution of which yields all the amplitudes. An operational method, however, avoids the routine of setting up the simultaneous equations and gives the amplitude in terms of the transfer function once a particular root is known. Thus using formula (4.21), with a unit step involved in the present case,

$$A_1 = \left[\frac{1}{s}\frac{\Theta_o}{\Theta_i}(s)(s - r_1)\right]_{s=r_1} = \left[\frac{-(1 - s/r_1)}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)}\right]_{s=r} \tag{6.6}$$

Either of these forms requires that the transfer function be known in analytical form. When using the conformal-mapping process, however, the data might be from experimental measurement, which means that the form of the function is not yet known. The problem, therefore, is to determine the amplitude in terms of measurements which can be made directly on the conformal map.

Note that (6.6) involves a numerator and denominator, both of which are zero upon substitution of r_1 for s. This ratio is indeterminate; therefore, the ratio will be expressed as the *limit* of the ratio as s *approaches* r_1.

$$A_1 = \lim_{s \to r_1} \frac{1 - s/r_1}{(\Theta_i/\Theta_o)(s)} = \lim_{\Delta s \to 0} \frac{-\Delta s/r_1}{(\Theta_i/\Theta_o)(s)} \tag{6.7}$$

The value of the $(\Theta_i/\Theta_o)(s)$ function can be expressed as

$$\left[\frac{\Theta_i}{\Theta_o}(s)\right]_{r_1+\Delta s} = \left[\frac{\Theta_i}{\Theta_o}(s)\right]_{r_1} + \left[\frac{d}{ds}\frac{\Theta_i}{\Theta_o}(s)\right]_{r_1} \Delta s \tag{6.8}$$

But the value of the function at r_1 is zero; therefore, substituting the remaining term of (6.8) into (6.7) gives

$$A_1 = \frac{-\Delta s/r_1}{\left[\dfrac{d}{ds}\dfrac{\Theta_i}{\Theta_o}(s)\right]_{r_1} \Delta s} \tag{6.9}$$

The incremental value Δs cancels, leaving a definite result, which involves the derivative of the function $(\Theta_i/\Theta_o)(s)$ with respect to s. This function, however, is just 1 plus the inverse loop function for a unity-feedback system:

$$\frac{d}{ds}\left[\frac{\Theta_i}{\Theta_o}(s)\right] = \frac{d}{ds}\left[1 + \frac{\Theta_e}{\Theta_o}(s)\right] = \frac{d}{ds}\left[\frac{\Theta_e}{\Theta_o}(s)\right] \tag{6.10}$$

A derivative is the limit of the ratio of a change in a function divided by the change in the variable. In this case the change in the function is a vector on the conformal map, and the change in the variable is denoted by the values of s on the curves (Fig. 6.9). For example, the change in the function between the lines of constant damping at the -1 point is a vector approximately equal to $0.95\underline{/95^\circ}$ with the corresponding change in s being $+0.5$. The change in the function between lines of constant frequency is approximately $0.80\underline{/185^\circ}$, for which the change in s is $j0.5$. The ratio, or derivative, in each case is approximately the same, $1.8\underline{/95^\circ}$.

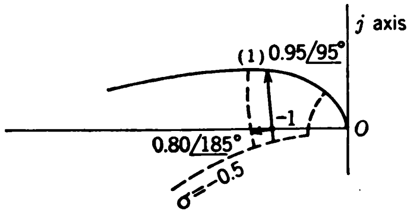


FIG. 6.9. Derivative of a function obtained from increments on a conformal map.

A more exact calculation can, of course, be made, but this one estimate is sufficient for the original purpose of this analysis, *i.e.*, checking whether the roots first determined from the conformal map dominate the transient response. Solve for A_1 from (6.9), using the value of the derivative obtained from Fig. 6.9 and $r_1 = 0.95\underline{/115^\circ}$. The first two terms of (6.5) are conjugate complex numbers, as pointed out in Art. 3.9; their sum is thus twice the real part of $A_1 e^{r_1 t}$

$$\begin{aligned} 2\Re A_1 e^{r_1 t} &= 2\Re \frac{-1}{0.95\underline{/115^\circ}} \frac{1}{1.8\underline{/95^\circ}} e^{-0.4t} e^{j0.85t} \\ &= 1.1e^{-0.4t} \cos (0.85t - 30^\circ) \qquad (6.11) \end{aligned}$$

A plot of this transient is shown in Fig. 6.10 by establishing the exponential decay envelope $1.1e^{-0.4t}$ and then multiplying by $-\cos (0.85t - 30^\circ)$.

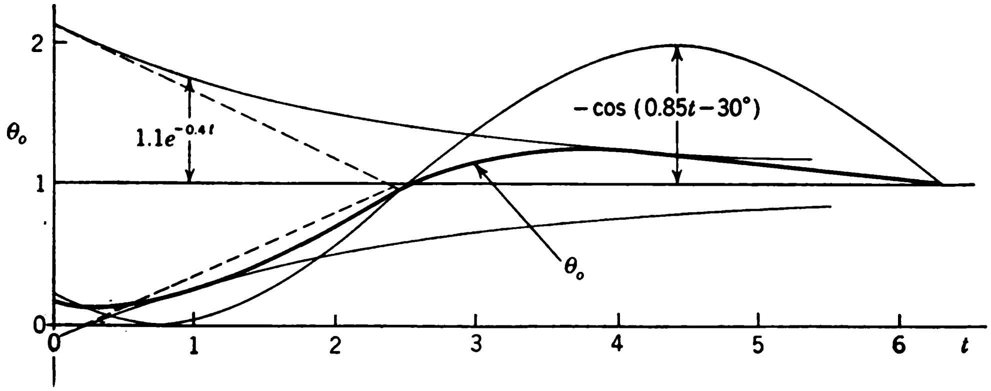


FIG. 6.10. Transient response corresponding to root determined in Fig. 6.8 and amplitude determined in Fig. 6.9.

This oscillation nearly satisfies the steady-state amplitude conditions required for the entire transient. The initial conditions are, of course, not satisfied because all the transient terms are needed to satisfy all these conditions. If the initial part of the response is not important, then the analysis need not be carried further. A complete solution, however, can be obtained, as indicated in the next article.

6.6. Complete Solution

Complex roots correspond to a quadratic factor in the transfer function. Subtracting the frequency-response curve of this factor from the system curve permits the other factors to be determined.

The frequency-response curve in Fig. 6.11 is that of the inverse system function if the -1 point is taken to be the origin. All the roots must be determined before the complete transient solution can be expressed. Supposedly the conformal map could be extended farther and the other roots found, even if they are real numbers. Such a plot would be difficult to sketch and probably inaccurate because of the extent of extrapolation from the frequency-response curve which would be required. A simpler procedure is to subtract the frequency-response characteristic known to be produced by the quadratic factor. Subtracting the frequency response of this factor leaves a remainder which is reasonably well matched by $1 + Ts$ in which $1/T = 4$.

The procedure for more complicated systems can be the same as for this system. Some complications arise, however, when the frequency-response curve bends around the -1 point. A conformal map built up from one portion of the curve then overlaps the map built up from the other side (Fig. 6.12). Two sets of complex roots are then present. A convenient procedure in this case is to make a rough estimate of one pair of roots and subtract the frequency response which would be due to this factor. The remaining curve can then be used as the boundary for curvilinear squares without overlap being present. The roots thus determined will not be exact, because the estimate of the first set of roots is not correct. The roots thus determined correspond to a quadratic factor whose frequency-response curve should be subtracted from the

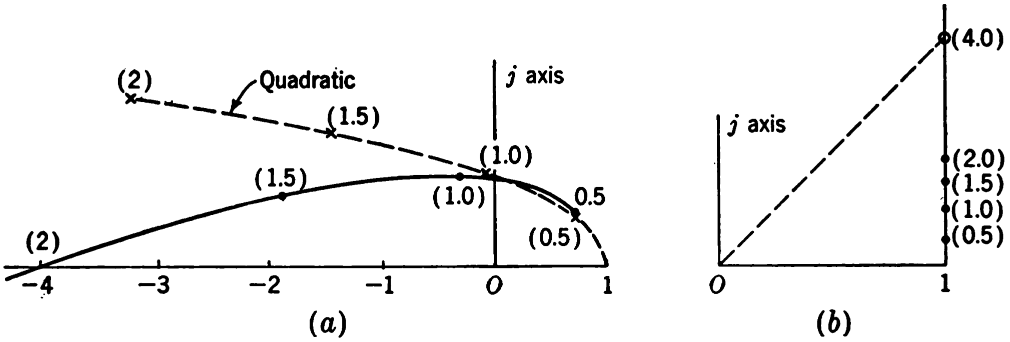


FIG. 6.11. Elimination of frequency response due to known quadratic factor: (a) plot of the inverse frequency-response curve and the quadratic; (b) the remaining frequency-response characteristic.

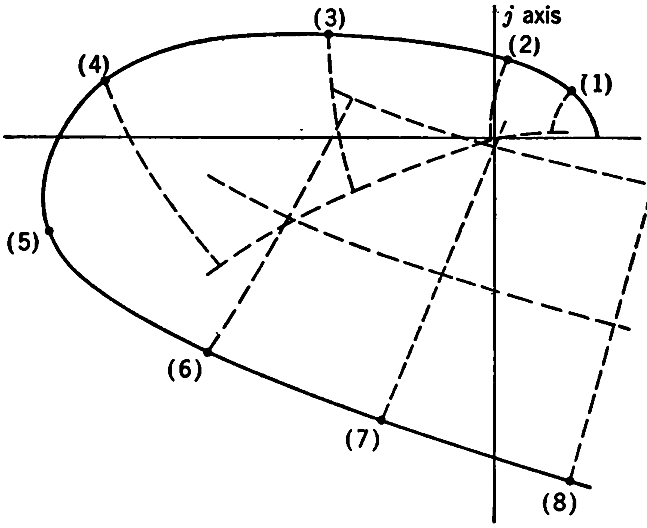


FIG. 6.12. A system having overlapping conformal grids.

original curve in order to make an improved estimate of the first pair of roots. This process can be repeated if necessary until consistent values of each pair of roots are obtained.

Removing a known factor from a vector plot involves subtracting phase angles and dividing vector lengths.

6.7. Conformal Mapping in Logarithmic Coordinates

The conformal-mapping technique used with vector plots can be used with logarithmic plots of attenuation versus phase angle if a consistent set of units is used.

The techniques of curve matching and conformal mapping can be used in combination if the conformal-mapping technique can be justified with logarithmic coordinates. The transfer function is treated as a vector quantity, to permit its logarithm to be taken:

$$\log \frac{\Theta_o}{\Theta_e}(s) = \log Re^{j\phi} = \log R + j\phi \qquad (6.12)$$

The real part is the logarithm to the base e of the magnitude of the function, for which the units are nepers. The j part is the angle of the vector expressed in radians. The units normally used for servo plots are decibels and degrees, each of which is an arbitrary choice. A consistent plot can be obtained, however, if the distance representing a magnitude ratio of e is the same distance as that for an angle of 1 rad; then 20 db, representing a factor of 10, is 2.3 nepers, for which the angle scale should be 2.3 rad, or 132°.

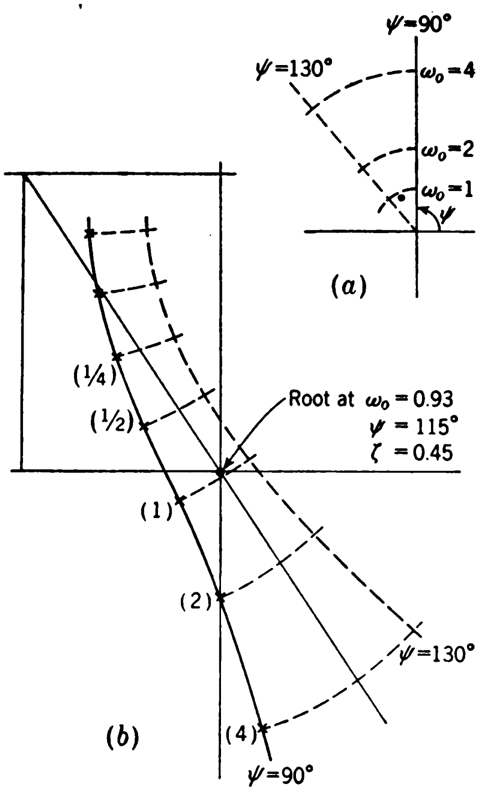


FIG. 6.13. Conformal mapping using logarithmic plots: (a) the curvilinear square grid in the s plane; (b) the conformal map of log $F(s)$.

The justification of a conformal map is the same as before; the increment in the function is proportional to the increment in the variable. Recall that the increments in the variable must be uniform in making a conformal map. In the vector plots these increments are a constant numerical change in σ and $j\omega$. For the logarithmic plots the increments in the variable are more convenient in logarithmic form:

$$\log s = \log \omega_0 e^{j\psi} = \log \omega_0 + j\psi \tag{6.13}$$

The increments in log ω_0 are a factor of 2, or 0.7 neper, and 40°, or 0.7 rad, as shown in Fig. 6.13*a*.

The conformal map is sketched in much the same manner as before, extending lines perpendicular to the frequency-response curve first, and then sketching lines roughly parallel to the curve separated by equal distances. Some adjustment is usually necessary to make squares throughout by curving the lines. The corresponding grid of lines in the *s* plane is shown in Fig. 6.13*b*.

The damping present is dependent upon the open-loop gain selected. For $K = 1$, as in Fig. 6.13*b*, the loop function is 180° in angle and unity in magnitude at $\omega_0 = 0.93$ and $\zeta = 0.45$. This result checks the value of r_1 obtained from the vector plot in Fig. 6.8.

6.8. An Introduction to the Root-locus Method

> *A root-locus plot in the s plane is the map of the negative real axis of the function plane. If the loop transfer function is known, the root-locus plot can be obtained more easily from it than by conformal mapping from a frequency-response curve.*

The conformal maps provide a picture of the effect of loop gain on the values of the dominant complex roots of a system. Thus in Fig. 6.8 a change in gain effectively changes the position of the -1 point in the vector plot and thus the value of the complex roots. In Fig. 6.13 a change in gain shifts the position of the 0-db line and thus the values of the complex roots. The maps thus provide a means of selecting loop gain to achieve the desired amount of damping in one mode of the transient response. If the frequency-response data are known only by experimental measurement, this mapping procedure is necessary. If the loop transfer function is known, however, calculations using complex values of *s* directly are preferable to calculations using only values of $j\omega$, with conformal mapping for the rest of the plot. The locus of the position of the roots with varying loop gain is the locus of values of the complex number *s* which make the loop transfer function have a phase shift of 180°. This locus proves to be established readily using the zeros and poles of the loop transfer function as pivot points for vectors. The loci of all roots are obtained. A knowledge of the conformal-mapping process used in this chapter is helpful but not necessary for making root-locus plots. Therefore, a separate introduction to the method is made in the next chapter in a way which does not depend upon frequency-response methods.

CHAPTER 7

THE ROOT-LOCUS METHOD

The root-locus method is a graphical method for finding the roots of the characteristic equation of the form $F(s) + 1 = 0$, in which $F(s)$ is in factored form. A single-loop control system has this form, where $F(s)$ is the direct transfer function around the loop for a system with negative feedback. The factors of $F(s)$ are the transfer functions of the components. If any component is known only by frequency-response data, then an approximation, like that discussed in Chap. 6, must be used to obtain its transfer function in analytic form. Then the zeros and poles of $F(s)$ are plotted in the s plane and used as the basis for sketching the locus of roots versus the loop gain as a variable. The method can also be applied to other problems if the equation can be put in the specified form, as illustrated in Chaps. 8 to 10.

7.1. Roots of the Characteristic Equation

> *The purpose of this section is to introduce the root-locus method in a way which does not depend upon the frequency-response method. This will be done by setting up the differential equation for a simple system and solving it by the classical method. The same problem will then be solved using the "transfer-function" concept. In either method the critical step is finding the "roots" of the "characteristic equation."*

In order to permit comparison of methods, the system to be considered is the same as that used in earlier chapters. The schematic diagram is shown in Fig. 4.1 and the block diagram in Fig. 5.1. For the present purposes, it is sufficient to state that the relationship between variables is described by the following differential equations:

$$\theta_e = \theta_i - \theta_o \tag{7.1}$$

$$K_rK_a\theta_e = i_1R_1 + L\frac{di_1}{dt} = R_1\left(i_1 + T_1\frac{di_1}{dt}\right) \tag{7.2}$$

$$K_g i = K_m\left(n + T_2\frac{dn}{dt}\right) \tag{7.3}$$

$$n = \frac{d\theta_o}{dt} \tag{7.4}$$

A single differential equation which relates the input θ_i to the output θ_o can be obtained by substituting from one equation into another. Specifically, substitute n from (7.4) into (7.3), then i_1 from (7.3) into (7.2), and finally θ_e from (7.2) into (7.1), with the result

$$\theta_i = \theta_o + \frac{1}{K}\frac{d\theta_o}{dt} + \frac{T_1 + T_2}{K}\frac{d^2\theta_o}{dt^2} + \frac{T_1T_2}{K}\frac{d^3\theta_o}{dt^3} \tag{7.5}$$

in which

$$K = K_rK_a\frac{1}{R_1}K_g\frac{1}{K_m} \tag{7.6}$$

If θ_i is given as a function of time, then part of the problem is to find the steady-state function of time for the output which will satisfy the differential equation (7.5). But, independent of this, the output can have a transient solution which makes the right-hand side of (7.5) equal to zero. These transient functions can be shown to be of the form e^{st}, with the value of s being determined by the constants of the system. Thus, assume that $\theta_o = Ae^{st}$ and substitute into (7.5) with the left-hand side equal to zero.

$$0 = Ae^{st} + \frac{1}{K}sAe^{st} + \frac{T_1 + T_2}{K}s^2Ae^{st} + \frac{T_1T_2}{K}s^3Ae^{st} \tag{7.7}$$

Note the very convenient result that the derivative of e^{st} is simply se^{st}. All terms in (7.7) have the same function of time, Ae^{st}, which may therefore be canceled, leaving an algebraic equation in s.

$$0 = 1 + \frac{1}{K}s + \frac{T_1 + T_2}{K}s^2 + \frac{T_1T_2}{K}s^3 \tag{7.8}$$

A value of s which satisfies this equation represents a function of time of the form e^{st} which will satisfy the differential equation with the input zero. The equation is called the *characteristic equation* of the system, and the values of s which satisfy it are called roots of that equation. This is a third-degree equation; hence there are three roots, r_1, r_2, and r_3. The transient response will thus be of the form

$$\theta_o = A_1e^{r_1t} + A_2e^{r_2t} + A_3e^{r_3t} \tag{7.9}$$

The values of the amplitudes A_1, A_2, and A_3 are determined by the initial conditions of the problem.

Consider now the same problem solved in the form of transfer functions. Make the following substitutions in (7.1) through (7.4):

$$\theta_e = \Theta_e e^{st} \qquad \theta_i = \Theta_i e^{st} \qquad \theta_o = \Theta_o e^{st}$$
$$i_1 = I_1e^{st} \qquad n = Ne^{st}$$

Cancel the common factor e^{st} throughout to obtain

$$\Theta_e = \Theta_i - \Theta_o \tag{7.10}$$

$$K_r K_a \Theta_e = R_1(1 + T_1 s) I_1 \tag{7.11}$$

$$K_g I_1 = K_m(1 + T_2 s) N \tag{7.12}$$

$$N = s\Theta_o \tag{7.13}$$

Express the ratio of the output variable to the input variable in each case:

$$\frac{I_1}{\Theta_e}(s) = \frac{K_r K_a / R_1}{1 + T_1 s} \tag{7.14}$$

$$\frac{N}{I_1}(s) = \frac{K_g / K_m}{1 + T_2 s} \tag{7.15}$$

$$\frac{\Theta_o}{N}(s) = \frac{1}{s} \tag{7.16}$$

Such functions, called transfer functions, describe the relationship between two variables and can be treated as algebraic quantities. The transfer function from θ_e to θ_o is thus the product of the individual transfer functions:

$$\frac{\Theta_o}{\Theta_e}(s) = \left[\frac{I_1}{\Theta_e}(s)\right]\left[\frac{N}{I_1}(s)\right]\left[\frac{\Theta_o}{N}(s)\right] = \frac{K}{(1 + T_1 s)(1 + T_2 s)s} \tag{7.17}$$

The transfer function from θ_i to θ_o can be expressed as

$$\frac{\Theta_o}{\Theta_i}(s) = \frac{(\Theta_o/\Theta_e)(s)}{1 + (\Theta_o/\Theta_e)(s)} \tag{7.18}$$

The algebra is simplified in this case, however, if the *inverse transfer function*, or ratio of input to output is considered.

$$\frac{\Theta_i}{\Theta_o}(s) = 1 + \frac{\Theta_e}{\Theta_o}(s) = 1 + \frac{(1 + T_1 s)(1 + T_2 s)s}{K} \tag{7.19}$$

Note that, if the ratio $(\Theta_i/\Theta_o)(s)$ is specified to be zero, then (7.19) is the same as the characteristic equation obtained by the classical method in (7.8). This result should not be surprising because exactly the same combination of equations is obtained in either case, even though in the classical method s did not replace d/dt until the substitution of an assumed solution of the form e^{st} in the final differential equation.

A significant difference in concept is present, however. The transfer functions of the individual components can still be identified in (7.19), whereas, in the form (7.8) in which (7.19) has been expanded into a general polynomial, the coefficients involve a combination of parameters such as $(T_1 + T_2)/K$. In a control-system design, the effect of changing the basic parameters, such as T_1 and K, must be known. Therefore, any method which determines the roots should do so in a way which preserves the effect of these parameters. Note also that (7.19) is simpler in form than (7.8) in that the factors of $(\Theta_e/\Theta_o)(s)$ are already known. This result suggests that the roots of the characteristic equation be found by making

$$\frac{\Theta_e}{\Theta_o}(s) = \frac{(1 + T_1s)(1 + T_2s)s}{K} = -1 \tag{7.20}$$

7.2. Vector Representation of the Open-loop Function

The problem of finding the values of s which make the open-loop transfer function equal to -1, and of finding them in a way which preserves the effect of each component, suggests that each term be treated as a vector. These vectors are shown to start from fixed points in the complex s plane and end at the s point being considered.

Consider first the frequency response of the function in (7.20) in vector form (Fig. 7.1). For a sinusoidal signal of frequency ω, the open-loop transfer function is found by substituting $s = j\omega$. Each term in (7.20) is shown in Fig. 7.1 as a vector. As the value of ω is varied, the heads of all the vectors shift together. This plot suggests that all vectors should be made to end at the s point. This result can be achieved by factoring out the coefficient of s from each term in (7.20). The same is true, of course, for the direct transfer function, which is usually considered.

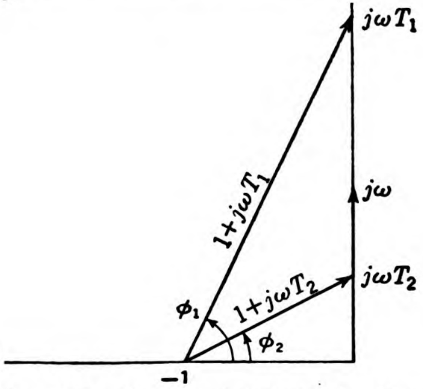


FIG. 7.1. Vector involved in the frequency response of the function $K/(1 + T_1s)(1 + T_2s)s$.

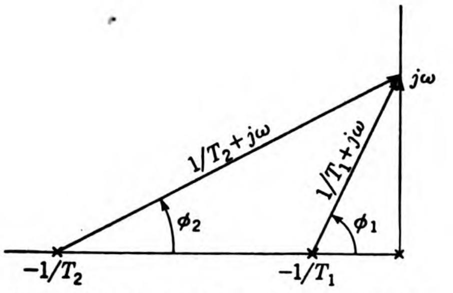


FIG. 7.2. Vectors having a common $j\omega$ point.

$$\frac{\Theta_o}{\Theta_e}(s) = \frac{K}{T_1[(1/T_1) + s]T_2[(1/T_2) + s]s} = -1 \tag{7.21}$$

The diagram for computing the frequency response of the transfer function of (7.21) is then as shown in Fig. 7.2.

The value of ω may now be allowed to assume various values along the vertical axis, and the heads of all vectors follow the s point with the tails fixed at 0, $-1/T_1$, and $-1/T_2$. These fixed points are denoted by crosses in Fig. 7.2.

In seeking to find roots, however, the value of s must be allowed to have a complex value $\sigma + j\omega$ rather than being limited to $j\omega$. The vector diagram of Fig. 7.2 is then altered merely by shifting the value of s off the $j\omega$ axis to a complex value, as shown in Fig. 7.3.

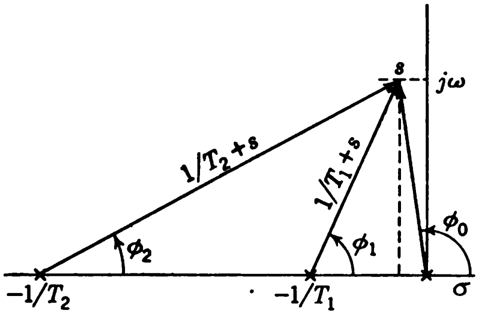


FIG. 7.3. Vector using a complex value of s.

This diagram provides a very rapid means of estimating the value of the transfer function for *complex values of* s. Each term of the function is treated as a vector; the product of terms is thus a vector whose angle is the sum of the angles and whose magnitude is the product of the magnitudes. The *condition for roots* expressed in (7.21) can be thus considered to be *two separate conditions:*

$$\phi_0 + \phi_1 + \phi_2 = 180° \tag{7.22}$$

$$\frac{K}{T_1|(1/T_1) + s|T_2|(1/T_2) + s||s|} = 1 \tag{7.23}$$

The condition (7.22) that the *sum of the phase angles be* 180° is more easily considered first, and an entire locus of values called the locus of roots will be shown to satisfy it. The second condition (7.23) is that the absolute magnitude of the loop transfer function be unity. The magnitude condition depends upon the static gain K and is considered later.

7.3. Sketching the Locus of Roots

> *The locus of roots is the path of all values of s for which the sum of the angles of the vectors from the poles and zeros of the open-loop function is equal to* 180°. *Special parts of the locus, such as intervals along the real axis and asymptotes for large values of s, are located by inspection, and the remainder of the locus is then sketched roughly.*

For values of s along the real axis, the angle of any vector of the given function is either 0° or 180°, as can be seen in Fig. 7.3. To the right of the origin all the angles are 0°, and so the locus cannot lie anywhere along that interval, since the total of the angles must be 180°. Between $-1/T_1$ and $-1/T_2$, two vectors are 180° and the third is still zero, giving a total of 360°, and hence the locus cannot exist in this interval. The locus does exist to the left of the $-1/T_2$ point, because all three vectors are then 180° and so the sum of the angles is 540°, or, subtracting 360° for a complete revolution, 180°.

The locus must somewhere break away from the real axis in order to move out to other parts of the locus in the complex plane. Such a *breakaway* point can be found by considering the s point to be just above the real axis, as shown in Fig. 7.4. All the angles shown in this figure are small and can be expressed in radians simply by the ratio of the vertical height to the proper base length. Thus, the angle by which the s vector is now less than 180° is $\Delta\omega/x_0$. Similarly, the angles of the other vectors are now $\Delta\omega/x_1$ and $\Delta\omega/x_2$ instead of 0. The sum of the angles is then still 180° if the angle by which the s vector is reduced is equal to the sum of the angles by which the other vectors are increased.

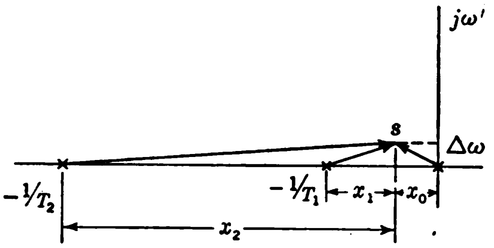


FIG. 7.4. Determination of breakaway points along the real axis.

$$\frac{\Delta\omega}{x_0} = \frac{\Delta\omega}{x_1} + \frac{\Delta\omega}{x_2} \tag{7.24}$$

brk pt

Note that the vertical height cancels, so that the breakaway point is determined simply by the reciprocals of the base lengths.

Another special case to consider arises when the magnitude of s is very large; then the angles of all the vectors are essentially equal, as shown in Fig. 7.5. If all the angles are equal, they must all be either ±60° or 180° in order that their sum be 180°. Thus, two branches of the locus must approach asymptotes which are at an angle of 60° with respect to the positive real axis, and the third branch is along the negative real axis.

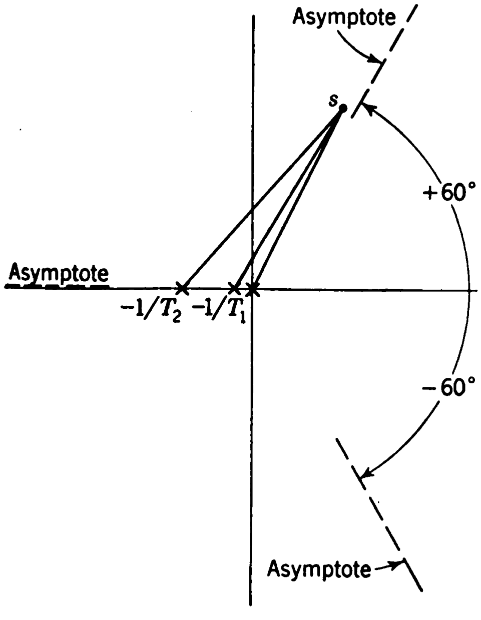


FIG. 7.5. Asymptotic limits of the locus for very large values of s.

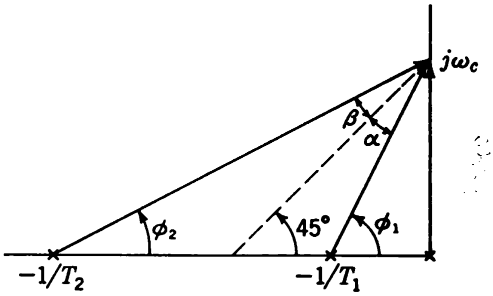


FIG. 7.6. Determination of crossover point of locus on the $j\omega$ axis.

The point at which the locus crosses the $j\omega$ axis is of particular significance and is also easy to locate, as shown in Fig. 7.6. The angle of the s vector is 90°; hence the crossing point must be such that the sum of the angles of the other two vectors is 90°. This point can be located in approximate position by sketching a 45° line from a point between the $-1/T_1$ and $-1/T_2$ points to the $j\omega$ axis. The angle ϕ_1 is then surely greater than 45° and the angle ϕ_2 less than 45°. The crossing point is correct if the angle α by which ϕ_1 is greater than 45° is equal to the angle β by which ϕ_2 is less than 45°.

The remainder of the locus is then sketched from the breakaway point on the real axis through the crossover points into the right half plane, finally approaching the 60° asymptotes (Fig. 7.7). Note that the plot is symmetrical about the real axis, because all angles on the lower part of the locus are equal in value but opposite in sign to those for the corresponding point on the upper part of the locus. The total angle $-180°$ is the same as 180° or $180° \pm n360°$, in which n is an integer, because the angle condition acts merely to make the transfer function negative.

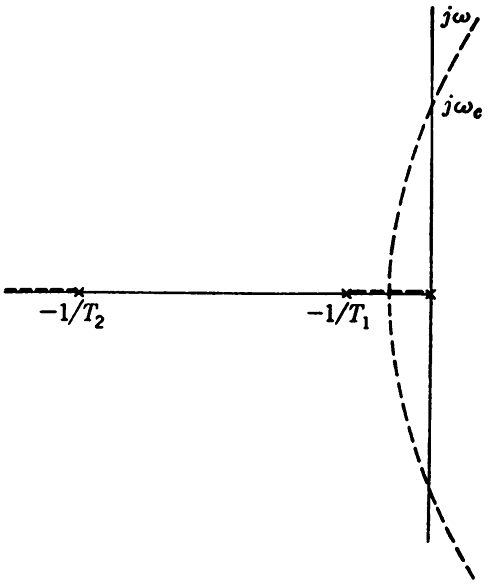


FIG. 7.7. Sketch of the complete locus from sections determined in Figs. 7.4 to 7.6.

7.4. Effect of Gain on the Position of the Roots

The roots lie along the locus at a position determined by the choice of loop gain. The roots start at the poles of the open-loop function for zero gain and move out along the locus for increasing gain.

The locus is obtained by satisfying only the condition that the total phase-shift angle of the open-loop transfer function be 180°. The second condition for an s point to be a root of the characteristic equation is that the magnitude of this function be unity.

$$\frac{K}{|1 + T_1 s||1 + T_2 s||s|} = 1 \tag{7.25}$$

Inspection of this function shows that, for very small values of gain K, the function can be near unity only if one of the denominator terms in (7.25) is also very small. This condition is satisfied for values of s near 0, $-1/T_1$, and $-1/T_2$. In the limit, gain K is zero, and the roots for

$K = 0$ must be exactly at these points in order that for each root a denominator term be zero. These points are the *poles* of the open-loop transfer function. *A pole is a value of s which makes a function infinite.* For increasing ~~gain,~~ the roots thus emerge from the poles and move out along the locus, as shown in Fig. 7.8.

The positions of the roots for successively higher values of gain are denoted in Fig. 7.8 by K_1, K_2, K_3, and K_4. The corresponding transient responses of the system for a step input are shown in Fig. 7.9. Gain K_1 produces the case of critical damping, in which the system just fails to overshoot. Increase in gain to K_2 causes an oscillation to appear, which increases in frequency with further increase in gain. The damping is gradually reduced until finally at K_3 a resonant system is obtained, in which an oscillation will maintain itself. Still further increase in gain to K_4 causes the amplitude of the oscillation to build up, the rate depending upon how much the gain has been raised above the value which produces the resonant system. The effect of the real root to the left of $-1/T_2$ will be shown in Art. 7.9 to be small.

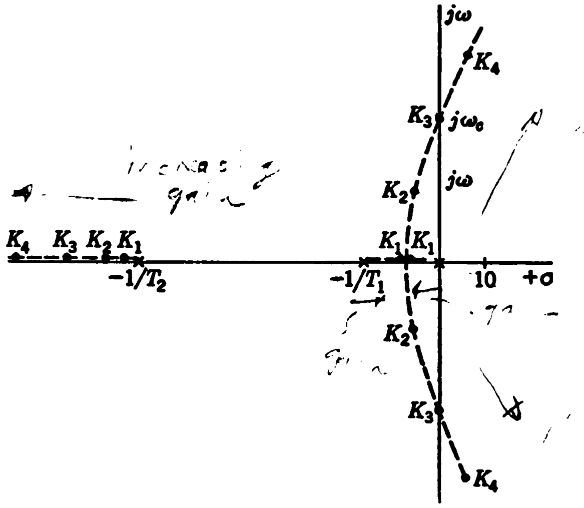


FIG. 7.8. Effect of loop gain on the position of the roots along the locus.

The estimate of the value of gain needed to put the roots at any particular position along the locus can be made by eye to an accuracy of about 20 per cent. The value of $|s|$ itself depends upon the scale of the plot. In Fig. 7.8 the length for 10 is shown to indicate the scale; hence the length of the s vector for the point K_2, say, would be estimated as about 15. A term such as $1 + T_1s$ is estimated as the ratio of two line lengths: $|1/T_1 + s|/|1/T_1|$. The length $1/T_1 + s$ from the pole $-1/T_1$ to the root s is about 20 per cent longer than the length $1/T_1$ from the pole $-1/T_1$ to the origin. Therefore the term $1 + T_1s$ is estimated as 1.2. Note that the scale of the drawing does not affect the value of such a ratio. The term $1 + T_2s$ is estimated similarly to be near 1, since

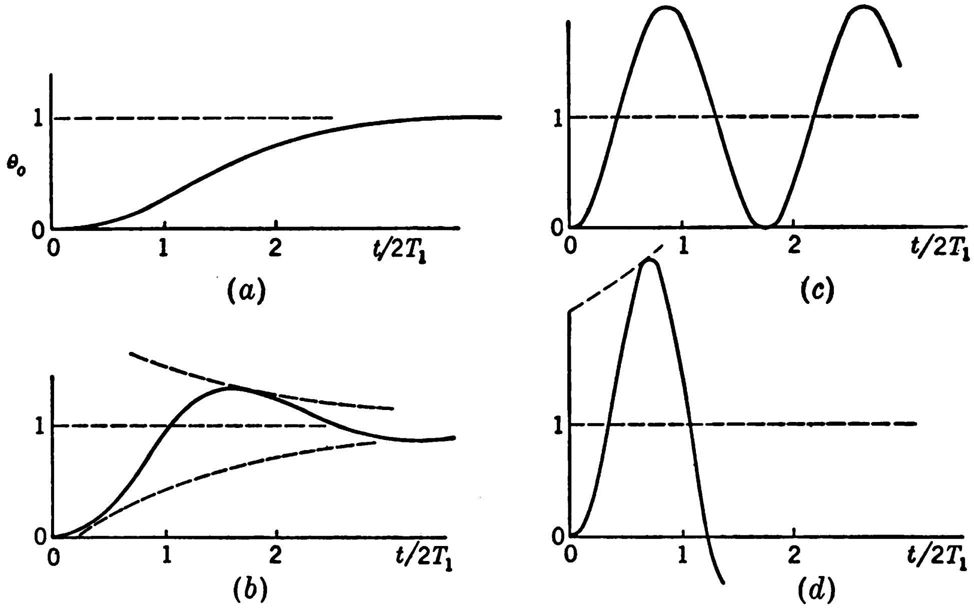


FIG. 7.9. Transient responses corresponding to positions of the pair of roots shown in Fig. 7.8, neglecting the real root to the left of $-1/T_2$: (*a*) K_1, critically damped case; (*b*) K_2, oscillation with 0.5 damping ratio; (*c*) K_3, continuous oscillation at constant amplitude; (*d*) K_4, oscillation which builds up exponentially.

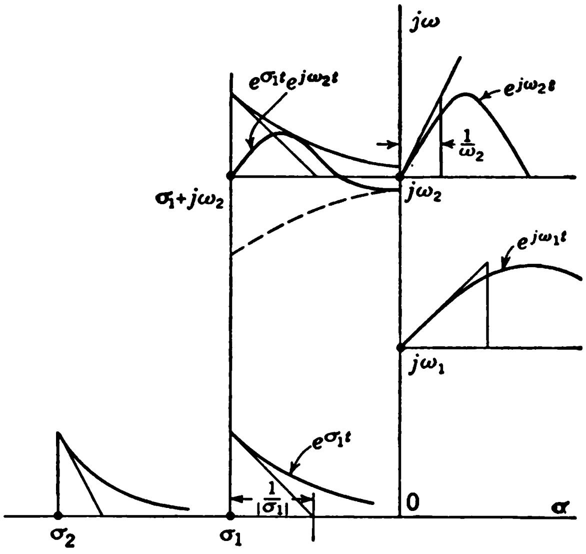


FIG. 7.10. Transients corresponding to various root positions.

the pole $-1/T_2$ is about equidistant from the root s and the origin. The gain K_2 then satisfies (7.25) by being equal to

$$(1 + T_1 s)(1 + T_2 s)s = 1.2 \times 1 \times 15 = 18$$

A more general picture of the significance of a root is shown in Fig. 7.10. The real root σ_1 is negative and has a time constant of decay of $1/|\sigma_1|$ as shown. Similarly a pair of roots at $j\omega_2$ corresponds to an undamped oscillation in which the time for 1 rad of oscillation is $1/\omega_2$. In general, for a complex root $\sigma_1 + j\omega_2$, the transient is a product of two curves $e^{\sigma_1 t}$ and $e^{j\omega_2 t}$, as shown. The other curves show the slower rate of oscillation of ω_1 compared to ω_2 and the more rapid decay for σ_2 than for σ_1.

The determination of the roots of a system is not the end of the analysis of that problem; the form of the closed-loop function and the transient response are usually desired. Rather than discuss these topics immediately, however, an example will be considered which involves a lead term and repeated poles.

7.5. The Effect of a Lead Term

> *A lead term introduces a term in the numerator of the direct transfer function; the zero is identified by a circle on the plot, and the angle of this vector is subtracted from the angles contributed by the denominator terms.*

poles

Consider the system shown in Fig. 7.11. Such a system can arise if the motor has negligible damping; the voltage impressed causes a proportional acceleration of the output regardless of speed. The signal feedback, however, is the sum of the position of the output and T times the velocity of the output as measured by a tachometer. The difference between the input signal θ_i and the feedback signal θ_f is now the signal to the amplifier θ_a. Note that this signal is no longer the error θ_e, which has been previously defined as being the difference between the input and the output. The roots of the system are again the values of s which make the transfer function around the loop equal to -1.

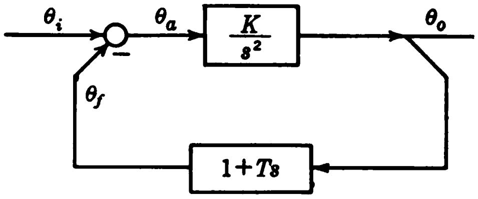


FIG. 7.11. A system involving multiple poles and a zero in the open-loop transfer function.

$$\frac{K(1 + Ts)}{s^2} = -1 \tag{7.26}$$

A new effect is now present; the term $1 + Ts$ is in the numerator rather than the denominator, making $-1/T$ a *zero*. *The value of s which makes a function zero is called a zero of that function.* In setting up the plot for the locus of roots, this zero $-1/T$ will be identified by a circle rather than a cross, as shown in Fig. 7.12.

The double pole at the origin, due to the presence of $1/s^2$ in (7.26), is denoted by two closely spaced crosses, with the understanding, of course, that both are actually right at the origin. The phase shift due to the term $1 + Ts$ in the function in (7.26) is a lead rather than a lag and hence is subtracted in finding the net lag of the loop transfer function. The phase shift of the function for the point s is then $2\phi_O - \phi_T$, and the condition that s be a root of the function is

$$2\phi_O - \phi_T = 180° \tag{7.27}$$

In sketching the locus, one might well check the intervals along the real axis first. Note that the locus does not exist between the origin and $-1/T$, since the two poles at the origin contribute a total of 360° along this interval. The locus does exist to the left of $-1/T$, however. Near the origin, the locus is along the $j\omega$ axis because each pole contributes 90° phase shift and the phase lead due to the zero at $-1/T$ is negligible. This portion of the locus arises for very low values of gain, as can be seen by inspection of (7.26) when both s and K have small values. As the gain of the system is increased, however, the magnitude of s must increase, and so the roots move out from the origin as noted by the arrowheads in Fig. 7.12.

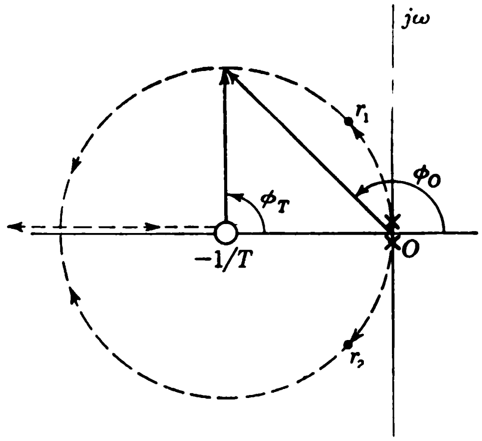


FIG. 7.12. Root-locus plot for the system shown in Fig. 7.11.

As the root r_1 moves upward, the lead from the zero at $-1/T$ increases. The lag from the poles at the origin must also increase in order to maintain the net angle 180° as specified in (7.27). This causes the locus to curve to the left as shown by the dashed lines. Looking for an intersection with a vertical line through $-1/T$, one notes that the 90° lead due to the vector $1/T + s$ can be balanced by making the phase angle of the vector from each pole at the origin equal to 135°, as shown. From this point one guesses that the locus curves around and enters the real axis. The point of entry into the real axis must be $-2/T$, using the small-angle reasoning illustrated for the previous system in (7.24). The locus thus determined can be shown to be a circle about the point $-1/T$ as a center.

In computing the loop gain needed for a particular point on the locus, the magnitude of s^2 is divided by the magnitude of $1 + Ts$ rather than being multiplied by it.

$$K = \frac{|s|^2}{|1 + Ts|} \tag{7.28}$$

Fortunately, these two examples involve all the effects which must be considered in a single-loop system: zeros and poles of the loop transfer function and multiple poles or zeros. Thus one can proceed with the solution of these two problems, being now aware of the effects to be encountered in any single-loop system.

7.6. Aids in Sketching Root-locus Plots

A root-locus plot can be sketched by using special cases, such as intervals along the real axis, asymptotes for large values of s, and modification of basic patterns. A summary of all aids, including some based on the theory of later chapters, is given in Appendix B.

The simplest parts of a root-locus plot to establish are the intervals along the real axis These intervals have an odd number of zeros and poles to the right of the interval. The breakaway points from the real axis depend upon the reciprocals of the base lengths to the zeros and poles (Fig. 7.4). For very large values of s, all angles are approximately equal, and therefore asymptotes of the locus are at $(180° \pm n360°)/m$, in which n is an integer and m is the excess of poles over zeros (Fig. 7.5). Patterns such as that in Fig. 7.12 are basic ones which appear in many plots.

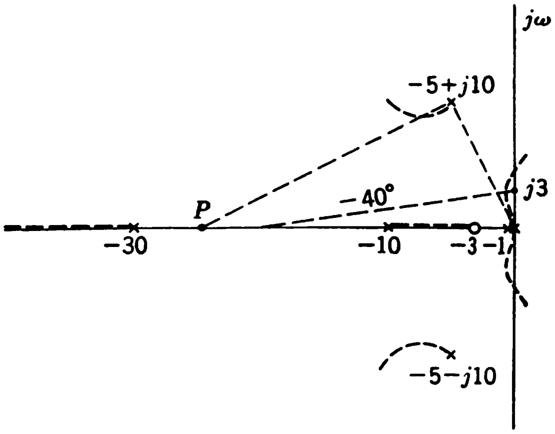


FIG. 7.13. Root-locus plot for poles distant from the origin.

Additional aids will be illustrated by a single example which includes the important problem of the choice of scale factor for the plot. Consider the case of a system whose loop transfer function is

$$F(s) = \underbrace{\frac{K \times 125}{s(1+s)(1+s/10)(s^2+10s+125)}}_{\text{Basic system}} \; \underbrace{\frac{1+s/3}{1+s/30}}_{\text{Lead net}} \; \underbrace{\frac{1+s/0.3}{1+s/0.03}}_{\text{Integral net}} \qquad (7.29)$$

The first plot is made with a scale such that the most extreme pole at −30 can be included, as shown in Fig. 7.13. The terms involved in the integral network are approximated as being equal to 10, because the measurements to the pole at 0.03 and the zero at 0.3 would be almost equal for s with a value of 3 or larger. The locus intervals along the real axis are located as for the previous problem, and the final asymptotes for large values of s are at $\pm 180°/5$, $\pm 540°/5$, and 180°. The direction in which the locus emerges from the pole at $-5 + j10$ can be found by considering an s point very close to this pole. The angles of all vectors to this s point except the vector $s + 5 - j10$ can be measured on the plot. The s point is then selected such that the angle of the vector $s + 5 - j10$ brings the total angle up to 180°. The dashed loci shown are then sketched to curve toward their final asymptotes. Additional aids in this sketching process are given in Appendix B.

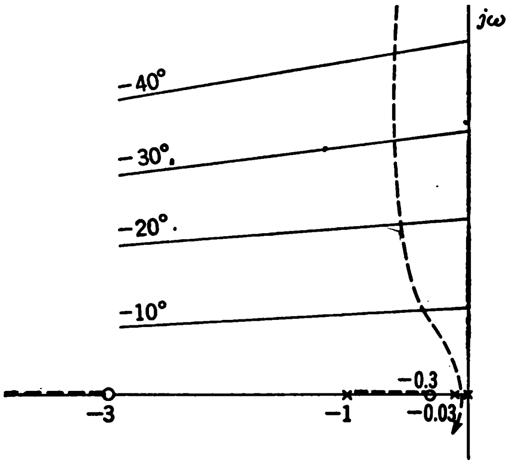


FIG. 7.14. Root-locus plot for region close to the origin.

The plot within the region where s is equal to or less than 3 requires a new scale, as shown in Fig. 7.14. The poles beyond $|s| = 3$ cannot be shown, but their effect must be considered. The magnitude of the terms in (7.29) such as $1 + s/30$ is about 1, but the phase angle measured at $j3$ in Fig. 7.13 due to poles beyond $|s| = 3$ is 40°. The effect of the complex poles on this phase shift at this point is shown in Appendix B to be almost equal to that of two poles at the point P in Fig. 7.13. The effect of the four poles on the axis is almost equal to that of four poles at -20 for any value of s within a radius of 3. The function in (7.29) can then be approximated, for s equal or less than 3, as

$$F(s) = \frac{K(1 + s/3)(1 + s/0.3)}{s(1 + s)(1 + s/20)^4(1 + s/0.03)} \tag{7.30}$$

The phase shift due to the four poles at -20 is marked on Fig. 7.14 by straight lines which would converge at -20. The locus is then sketched such that the net shift due to the poles and zeros on the plot is less than 180° by the angle marked for the four poles.

Additional aids in sketching plots are given in Appendix B rather than being explained throughout the book in connection with individual problems. After a plot has been sketched, the parts of the locus of interest are then checked, as explained in the next article.

7.7. Aids in Calculating Root-locus Plots

All the vectors on the root-locus plot have their heads at the s point and their tails at the zeros and poles. The angles and the vector *lengths can therefore all be measured at this point. A disk pivoted at the s point can be rotated through each of the angles in succession to obtain their sum. A logarithmic scale permits the multiplication of vector lengths by the addition of logarithms. A plastic device incorporating these features is called a Spirule and is described in Appendix C.*

A root-locus plot should first be sketched by use of the special cases described in the previous sections. A compilation of such aids in sketching is given in Appendix B. Parts of the locus of interest can then be checked by selecting a trial point and measuring the total angle of the vectors. Note that all the angles are present at the *s* point, as shown in Fig. 7.15. A protractor could be placed at this point, and the numerical value of the angles read and added. A faster procedure, however, is to pivot a piece of vellum or clear plastic at the *s* point and rotate the vellum through each of the angles in succession to obtain their sum. Pivoting the vellum can be achieved by simply pressing the point of a pencil on the vellum at the *s* point. Rotation can be achieved by putting the sharpened point of an eraser on the vellum over a pole of the plot and pushing the vellum with the eraser until the eraser is along a horizontal line to the left of the *s* point. An arrow marked on the vellum which is originally aligned horizontally to the right should be horizontal to the left after all the angles have been added if the sum is to be 180°. Rotation for zeros should, of course, be opposite in direction from that for poles. Two trial points in a given region are usually sufficient to establish a point on the locus by interpolation. An exception is the case of two nearly equal roots; for this case any point in the region gives nearly 180° total phase angle (see Appendix B). The position of the roots, in this case, is quite indefinite but is also not critical in computing either transient response or the frequency response of the complete system.

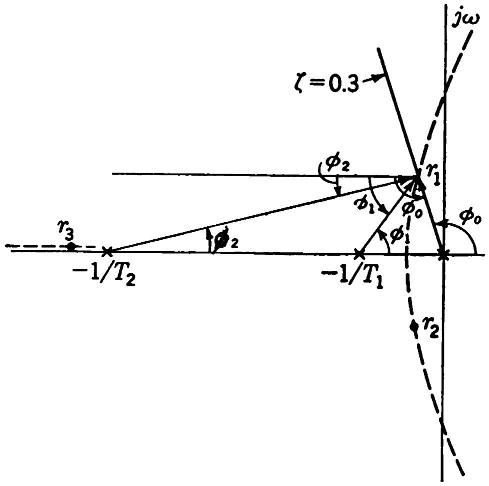


FIG. 7.15. Phase angles measured at the root.

The vector lengths could be measured numerically and multiplied on a slide rule. A logarithmic scale, like that shown in Fig. 7.16, however, permits the decibel value of each length to be read directly, and these values can be added to obtain the total decibel value of the product. If the scale is placed at the origin, then the decibel values of all the base lengths, such as $1/T_1$ in Fig. 7.15, can be read and subtracted from the decibel value obtained at the s point to obtain the decibel value of the gain factor K. This value is subject to a correction for the s vector itself, because unity on the scale will probably not correspond to the unit length used on the plot. The correction, in this case, is $20 \log_{10} x$, in which x is the numerical value on the plot corresponding to a unit length on the scale.

The above aids in calculation, and more, have been incorporated into a plastic device called the "Spirule" (Fig. 7.17). The Spirule consists

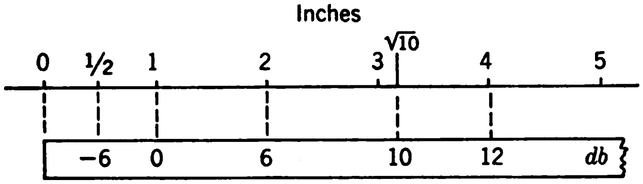


Fig. 7.16. Logarithmic scale for product of vector lengths.

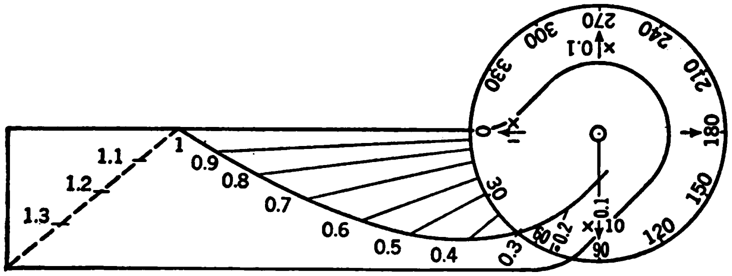


Fig. 7.17. The Spirule.

of a disk and arm held together with a light friction fit by an eyelet. A pin, mounted in the center of the eyelet with a clear plastic plug, serves as a pivot for all calculations at the s point. The arm is rotated, with respect to the disk, through each of the vector angles in succession to obtain their sum. A logarithmic spiral curve on the arm permits the logarithm of a length to be obtained as an angle, so that addition of such angles corresponds to adding logarithms. A complete description of the Spirule is given in Appendix C.

7.8. The Closed-loop Function

The roots of the characteristic equation are the poles of the closed-loop transfer function. Knowing the roots thus permits either the frequency response or the transient response to be obtained.

The original expression (7.19) for the ratio of the input to the output of the system in Fig. 5.1 is

$$\frac{\Theta_i}{\Theta_o}(s) = 1 + \frac{\Theta_e}{\Theta_o}(s) = 1 + \frac{(1 + T_1 s)(1 + T_2 s)s}{K} \tag{7.31}$$

The root-locus plot, shown in Fig. 7.15, determines three roots which make $(\Theta_i/\Theta_o)(s) = 0$; the root r_1 shown, its conjugate r_2, and r_3 along the branch to the left of $-1/T_2$. The function in (7.31) can be rewritten in terms of the roots as in

$$\frac{\Theta_i}{\Theta_o}(s) = C(s - r_1)(s - r_2)(s - r_3) \tag{7.32}$$

The form in (7.32) makes the function zero for values of s equal to any one of the roots. The constant C is yet to be specified. First let the factors involving s be put in a form such that they are equal to 1 when $s = 0$.

$$\frac{\Theta_i}{\Theta_o}(s) = C'\left(1 - \frac{s}{r_1}\right)\left(1 - \frac{s}{r_2}\right)\left(1 - \frac{s}{r_3}\right) \tag{7.33}$$

Note that function in (7.31) is 1 for $s = 0$; therefore this same function $(\Theta_i/\Theta_o)(s)$ in (7.33) must be 1 for $s = 0$, making $C' = 1$. A check on this form is provided by realizing that the function in (7.31) could be expanded into a cubic polynomial in s, as could the same function in (7.33). In general, four constants could be arbitrarily selected in (7.31), and four constants are also provided in (7.33) to permit a solution.

The direct transfer function is more frequently considered; therefore (7.33) would appear as in

$$\frac{\Theta_o}{\Theta_i}(s) = \frac{1}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)} \tag{7.34}$$

This result means that the closed-loop system shown in Fig. 5.1 can now be replaced by an equivalent single block, as shown in Fig. 7.18. This representation permits the closed loop to be treated now just exactly as a single component was treated before. Thus, the transient produced by a step input can be written by inspection by use of operational methods, as shown in the next topic, or if the new block is part of another loop, it can be treated as a component of that loop. The frequency response can be expressed in a vector form similar to Fig. 7.2.

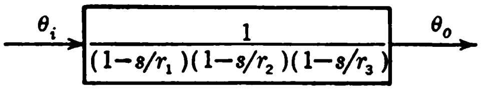


FIG. 7.18. Single-block equivalent of Fig. 5.1 made possible by determination of the roots.

The inverse transfer function between input and output[1] of the system shown in Fig. 7.11 is given by

$$\frac{\Theta_i}{\Theta_o}(s) = \frac{\Theta_a}{\Theta_o}(s) + \frac{\Theta_f}{\Theta_o}(s) = \frac{s^2}{K} + (1 + Ts) \tag{7.35}$$

---

[1] A transfer function $(\Theta_o/\Theta_i)(s)$ is said to be the transfer function *from* θ_i *to* θ_o because θ_i is the cause and θ_o the effect. When for convenience the inverse transfer function $(\Theta_i/\Theta_o)(s)$ is written, it is said to be *between* θ_i and θ_o to avoid implying that θ_o causes θ_i.

The values of s which make this function zero are the roots determined in the plot shown in Fig. 7.12 and are denoted as r_1 and r_2. The transfer function given in (7.35) can thus also be expressed as

$$\frac{\Theta_i}{\Theta_o}(s) = \left(1 - \frac{s}{r_1}\right)\left(1 - \frac{s}{r_2}\right) \tag{7.36}$$

Note that the value of this function is 1 for the case of $s = 0$, which checks the fact that the function in (7.35) is also 1 for $s = 0$.

The two examples thus far studied have involved only poles in the forward function and zeros in the feedback function. A more general system would have zeros and poles in both functions, as shown in Fig. 7.19.

The roots of the characteristic equation of this system are again the values of s which make the transfer function around the loop equal to -1.

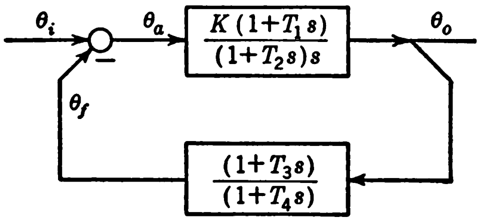


FIG. 7.19. A general single-loop system.

$$\frac{K(1 + T_1s)(1 + T_3s)}{(1 + T_2s)s(1 + T_4s)} = -1 \tag{7.37}$$

Note that no distinction is made between the forward function and the feedback function so far as finding the roots is concerned. In expressing the transfer function relating input to output, however, a distinction does arise.

The over-all inverse transfer function between input and output is the sum of the inverse function of the forward path and the direct function in the feedback path.

$$\frac{\Theta_i}{\Theta_o}(s) = \frac{\Theta_a}{\Theta_o}(s) + \frac{\Theta_f}{\Theta_o}(s) = \frac{(1 + T_2s)s}{K(1 + T_1s)} + \frac{1 + T_3s}{1 + T_4s} \tag{7.38}$$

The values of s which make this function zero are the roots of the characteristic equation of the system and may be determined from a root-locus plot which satisfies (7.37). Note, however, that the function in (7.38) is infinite for s equal to either $-1/T_1$ or $-1/T_4$. This fact must remain true in any form in which the function in (7.38) is rewritten. This condition is satisfied by keeping the factors $1 + T_1s$ and $1 + T_4s$ in the denominator; this is done in the familiar process of combining two fractions by making each have the same denominator. Thus (7.38) becomes

$$\frac{\Theta_i}{\Theta_o}(s) = \frac{[(1 + T_2s)s(1 + T_4s)/K] + (1 + T_3s)(1 + T_1s)}{(1 + T_1s)(1 + T_4s)} \tag{7.39}$$

The values of s which make the numerator zero are still the roots r_1,

r_2, and r_3; thus the function can be written in completely factored form as

$$\frac{\Theta_i}{\Theta_o}(s) = \frac{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)}{(1 + T_1s)(1 + T_4s)} \tag{7.40}$$

Again the d-c value of the function, obtained by substituting $s = 0$, is unity in the original equation (7.38) and in the final equation (7.40). The significant result of this algebraic manipulation is that the poles of the over-all inverse transfer function are $-1/T_1$ and $-1/T_4$.

The direct transfer function, or ratio of output to input, is more commonly used, however; in this function the values $-1/T_1$ and $-1/T_4$ are zeros. Speaking only in terms of *direct functions*, one can thus generalize to state that the *zeros of the over-all transfer function are the zeros of the forward function and the poles of the feedback function.*

7.9. Amplitude of Transient Terms

> *The transient response of a system is the sum of exponential terms, one for each root of the characteristic equation. The amplitude of each term is given directly in terms of vectors by using formulas developed in the Laplace transform or similar methods.*

A transient can be produced by almost any disturbance to a system. The system is described by its response to some standard disturbance, such as a sudden change of unit amplitude, known as a unit step. The steady-state change in the output of a system having unity feedback is then unity also. The transient is the sum of terms of the form Ae^{st}, one for each root of the characteristic equation. The response of the system shown in Fig. 5.1 is then

$$\theta_o = 1 + A_1e^{r_1t} + A_2e^{r_2t} + A_3e^{r_3t} \tag{7.41}$$

The amplitude of each term is given by a formula developed in operational calculus, or the Laplace transform theory (see Appendix A or Art. 4.9). Using the terminology of the Laplace transform, one would say that the differential equation relating input and output is *transformed* into an algebraic equation which states that the *transform* of the output is equal to the transform of the input times the transfer function between them. The transform of a step input can be shown to be $1/s$, and the transfer function $(\Theta_o/\Theta_i)(s)$ given in (7.34). The amplitude of a transient term in the output is obtained by substituting $s = r_3$ into the transform of the output with the factor $s - r_3$ omitted from the denominator:

$$A_3 = \left[\left(\frac{1}{s}\right)(-r_3)\left(\frac{-r_1}{s - r_1}\right)\left(\frac{-r_2}{s - r_2}\right)\right]_{s=r_3} \tag{7.42}$$

Note that the transfer-function terms are again represented as the ratio of terms and that the coefficient of s is 1 in each factor involving s.

The significant thing is that the terms thus obtained are all vector lengths on the root-locus plot, as shown in Fig. 7.20. The denominator terms are vectors from the other roots r_1 and r_2 to the given root r_3. The numerator terms are the vectors from the other roots to the origin. From symmetry, in this case, the phase angles of the vectors can be seen to cancel. The magnitudes can be estimated by eye just as in calculating gain on the root-locus plot. To maintain the comparison completely, the individual term will be estimated in the form of a ratio of line lengths, using the length to the origin as the base value, as in

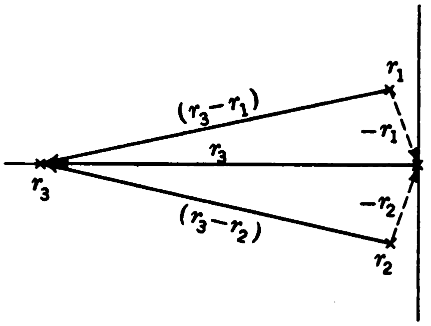


Fig. 7.20. Vector lengths involved in calculating the amplitude of the transient term $A_3e^{r_3t}$.

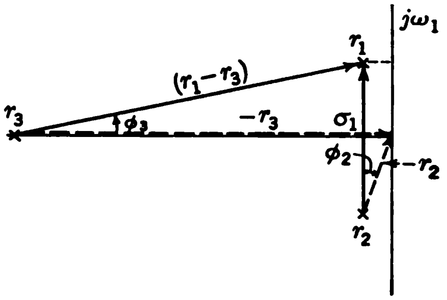


Fig. 7.21. Vectors involved in calculating the amplitude and phase angle of A_1.

$$\left[1-\frac{s}{r_1}\right]_{s=r_3} = \left[\frac{s-r_1}{-r_1}\right]_{s=r_3} = \frac{|r_3-r_1|}{|-r_1|} = 3 \tag{7.43}$$

The amplitude A_3 is thus approximately $\frac{1}{9}$, and the sign is negative owing to the negative sign of the product $(1/r_3)(-r_3)$ in (7.42).

Similarly, the amplitude A_1 is given by

$$A_1 = \left[\left(\frac{1}{s}\right)(-r_1)\left(\frac{-r_2}{s-r_2}\right)\left(\frac{-r_3}{s-r_3}\right)\right]_{s=r_1} \tag{7.44}$$

The corresponding vector lengths are shown in Fig. 7.21. The approximate value of A_1 is then

$$A_1 = (-1)\,\frac{|r_2|}{2\omega_2\underline{/\phi_2}}\,\frac{1}{0.9\underline{/\phi_3}} = \frac{-1}{2(0.8)}\,e^{-j45°} \tag{7.45}$$

The roots r_1 and r_2 are conjugate complex numbers, however, and so the two exponential transients involving them can be combined into a single cosine term. A simple means of effecting this conversion is to note that A_2 is the conjugate of A_1. At all values of time the j parts of the terms cancel and the real parts add, making the sum of the terms twice the real part of one of them:

$$
\begin{aligned}
A_1e^{r_1t} + A_2e^{r_2t} &= 2\Re(A_1e^{r_1t}) \\
&= 2\Re\,\frac{-1}{2(0.8)}\,e^{-j45°}e^{\sigma_1t}e^{j\omega_1t} \\
&= \frac{-1}{0.8}\,e^{\sigma_1t}\cos(\omega_1t - 45°) \qquad (7.46)
\end{aligned}
$$

Comparing this solution with the form which would be obtained for a pure quadratic, one notes that the damped sinusoid is changed in magnitude by the factor 1/0.9 and shifted in phase by ϕ_3 by the effect of the third root r_3. The complete solution for the response of the system is then

$$\theta_o = 1 - \frac{1}{9}e^{r_3t} - \left(\frac{1}{0.8}\right)e^{\sigma_1t}\cos(\omega_1t - 45°) \qquad (7.47)$$

Note that the effect of the third root becomes less and less for values of r_3 farther out the negative real axis. This is true not only because the transient term decays faster, but also because its initial amplitude is smaller.

7.10. Use of an Integral Network

The study thus far has been concerned with the determination of the output response for a given input. The error of the system can be noted as the difference between the two at any time. When seeking means of reducing the error, however, it is worthwhile to consider the transfer function from input to error.

Consider the system shown in Fig. 5.1 whose input-to-output transfer function is given in (7.34). The transfer function from input to error is given by

$$
\begin{aligned}
\left(\frac{\Theta_e}{\Theta_i}(s)\right) &= \left(\frac{\Theta_o}{\Theta_i}(s)\right)\left(\frac{\Theta_e}{\Theta_o}(s)\right) \\
&= \frac{1}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)}\,\frac{(1 + T_1s)(1 + T_2s)s}{K} \qquad (7.48)
\end{aligned}
$$

The steady-state error for an input having a constant velocity is simply $1/K$. This result is obtained by solving $(1/s)[(\Theta_e/\Theta_i)(s)]$ for the special case of $s = 0$. To decrease this error, the gain K must be increased, but stability requirements dictate that a circuit be added which will attenuate signals at the frequency range at which the system tends to be resonant. Such a circuit is the familiar integral net shown in Fig. 5.21*a*.

$$\frac{V_o}{V_i}(s) = \frac{R_2 + (1/Cs)}{R_1 + R_2 + (1/Cs)} = \frac{1 + R_2Cs}{1 + (R_1 + R_2)Cs} = \frac{1 + \alpha T_3s}{1 + T_3s} \qquad (7.49)$$

The attenuation of the network with the capacitor considered as a short circuit is $\alpha = R_2/(R_1 + R_2)$ and is often selected to be about $\frac{1}{10}$. This attenuation is achieved for the natural damped oscillation of the loop $e^{r_1 t}$ if $\alpha T_3|r_1|$ is about 5 or higher. The corresponding locus of roots is shown in Fig. 7.22.

The roots r_1', r_2', and r_3' are in essentially the same position as the corresponding part of the locus in Fig. 7.8. A new root r_4 appears and approaches the position of the zero $-1/\alpha T_3$ very closely. The phase lag at r_1' due to the integral network makes the damping of this root slightly less than it would otherwise be. The transfer function from input to output for this system is given by

$$\frac{\Theta_o}{\Theta_i}(s) = \frac{1}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)} \frac{1 + \alpha T_3 s}{1 - s/r_4} \tag{7.50}$$

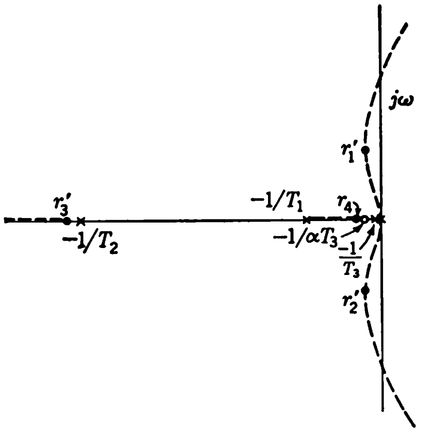


FIG. 7.22. Root-locus plot for basic system with an integral network added.

The new pole of this function at r_4 can be shown to have a transient of very small amplitude because the term $1 + \alpha T_3 s$ is small for $s = r_4$. All the other poles are nearly the same as those in (7.34), so that the behavior of the system between input and output is not appreciably affected by the addition of the integral network. The inverse transfer function $(\Theta_e/\Theta_o)(s)$, however, has the additional terms, due to the integral net and increased static gain of 10, of

$$\frac{\Theta_e'}{\Theta_e}(s) = \frac{1 + T_3 s}{(1 + T_3 s/10)10} = \frac{1 + T_3 s}{10 + T_3 s} \tag{7.51}$$

in which θ_e is the error before addition of the integral network and θ_e' is the error after addition of the integral network.

The significance of this transfer function in (7.51) can be more easily pictured if the numerator is divided by the denominator, giving

$$\frac{\Theta_e'}{\Theta_e}(s) = 1 - \frac{0.9}{1 + T_3 s/10} \tag{7.52}$$

Consider now that the input to the system had been a unit velocity started suddenly; the error can be determined in steps, as shown in Fig. 7.23. The error of the original system θ_e follows the input exactly at first, but as the output comes up to speed, the error levels off to the constant value $1/K$. All of this response is rapid, however, compared to the time constant $T_3/10$ of the additional transfer function due to the integral network. The new error θ_e' follows the original error at first but then decays, with the time constant $T_3/10$, to one-tenth the steady-state value previously required.

If the time delay $T_3/10$ in arriving at this reduced value of error is too long, then the value of capacitance used in the integral network must be reduced. This change in turn leads to more phase shift in the region of the root r'_1, and hence less damping, unless some compensating lead phase shift is introduced. The effect of a lead network is the next topic, although for simplicity it is considered with the original system.

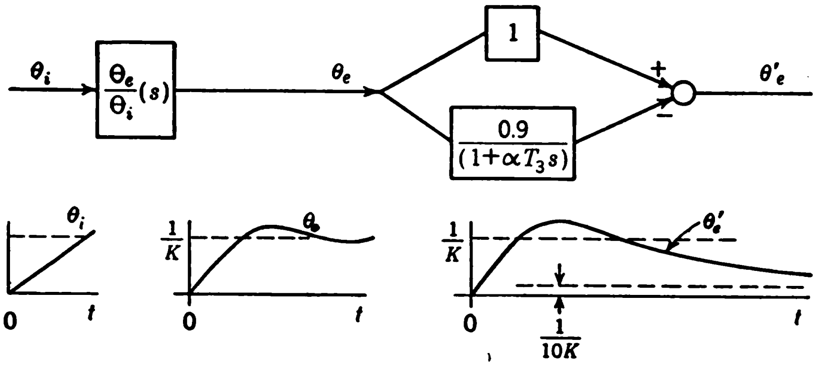


FIG. 7.23. Effect of integral network on error caused by sudden change in input velocity.

7.11. Use of Lead Network

> *A lead network can produce* 90° *phase lead for a root having a damping ratio of* 0.6. *The locus of the original system can thus be plotted for a total phase lag of* 270°. *The intersection of this locus with the radial line for* $\zeta = 0.6$ *thus determines the position of the root and the choice of the time constants in the lead network*[1].

The concept of the effect of a lead network described in the frequency-response method (Art. 5.11) is extended somewhat in dealing with complex values of roots. The circuit is shown in Fig. 5.23, and its transfer function is, from Eq. (5.29),

$$\frac{V_o}{V_i}(s) = \frac{\alpha(1 + Ts)}{1 + \alpha Ts} \qquad \alpha = \frac{R_2}{R_1 + R_2} \qquad T = R_1C_1 \qquad (7.53)$$

The characteristics of this function for complex values of s are conveniently plotted in the s plane as lines of constant phase angle, like the one shown in Fig. 7.24.

The locus of values of s for 90° phase shift proves to be a circle whose diameter is the interval along the negative real axis between the zero and the pole. This result can be shown by considering the dashed vectors from the zero and the pole to the point inscribed in the semicircle. The vector from the zero has the phase lead ϕ_1, and the vector from the pole has the lag ϕ_2; the net phase lead is the angle between the two vectors, as shown. For any point along the semicircle, the angle between the vectors will be 90°; hence this semicircle must be the locus for 90° phase lead of the network. The loci for the other angles of phase shift, such as 60° or 45°, prove to be circles also, with centers displaced vertically from the one shown, but the consideration of 90° phase lead happens to be particularly convenient.

---

[1] This direct solution for a lead network was suggested by W. E. Dickenson, now with International Business Machines at San Jose, Calif.

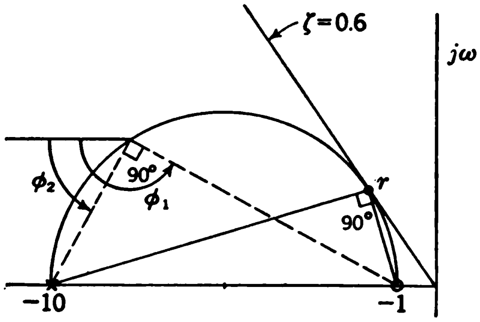


FIG. 7.24. Locus of values of s for which the phase shift of a lead network is 90°.

The radial line from the origin which is tangent to the circle at point r is the locus of points for which the damping ratio happens to be almost 0.6. If the criterion for the roots of the complete system is that the damping ratio shall be 0.6, then this figure shows the relative position of the zero and pole for most effective use of the lead network. In particular, the value of $1/T$ happens to be very nearly one-half the damping component of the root.

Since 90° phase lead can be attained for any value of s with the specified damping ratio of 0.6, the locus of values of s for which the rest of the system has a phase shift of 270° should be plotted. The point where this locus intersects the radial line for damping ratio of 0.6 will be the position of the complex roots of the system. This locus and the point of intersection are shown in Fig. 7.25. The locus starts vertically from the pole at $-1/T_1$ and then bends to the left. It approaches the asymptote shown whose horizontal position is determined by the condition that $x_2 = x_1 + x_0$. The reasoning here is that for very large values of s all the angles are nearly 90°, but the small angles by which they differ from 90° are proportional to the horizontal intercepts. The point of intersection with the line for damping ratio of 0.6 is then r, for which the angles can be checked by direct measurement, as by using the Spirule.

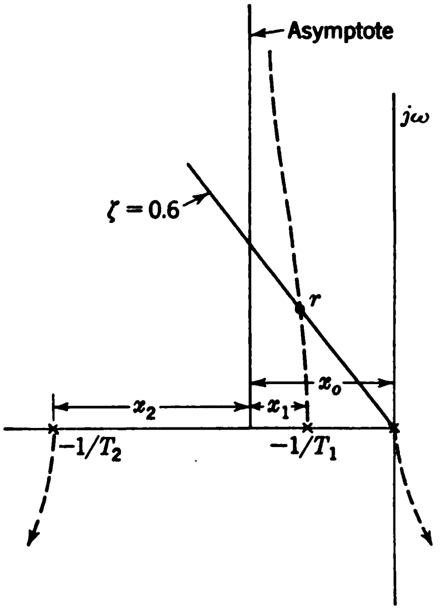


FIG. 7.25. Locus of values of s for which transfer function of basic system is 270°.

The damping of this root is σ_1; thus the zero of the lead network should be placed at $\sigma_1/2$ in order that the lead network may introduce 90° phase lead at point r.

The complete locus of roots of the system is as shown in Fig. 7.26. Note that the optimum value T_3 of the lead time constant for this system and this criterion is greater than the motor time constant T_1. If the value of the other time constant T_2 were only one-sixth of T_1, rather than one-fourth of T_1 as in Fig. 7.26, then the optimum value T_3 would be equal to T_1. For $T_2/T_1 < \frac{1}{6}$, T_3/T_1 should be less than 1 to obtain $\zeta = 0.6$.

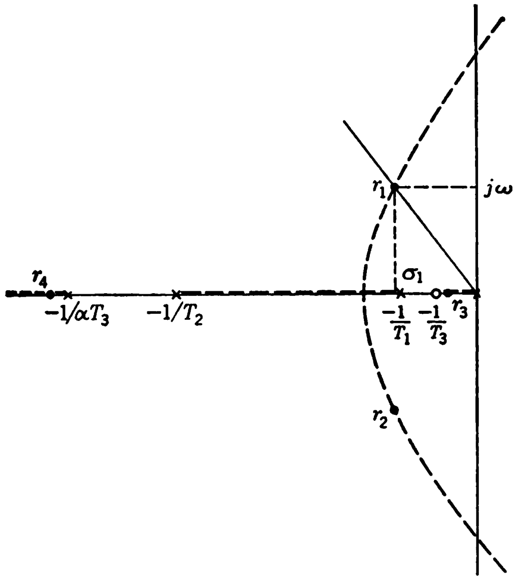


Fig. 7.26. Root-locus plot for basic system with the optimum lead network selected in Fig. 7.25.

In general, the system for which the lead circuit must be selected is more complicated than the cubic system above. The methods described here are limited in that the locus must always be plotted against loop gain as the parameter; Chap. 8 shows the technique for plotting the locus of roots versus C, the capacitance.

7.12. Comparison with the Frequency-response Method

The frequency-response method and the root-locus method both provide techniques for finding the response of a closed-loop system from its open-loop response in a way which suggests the changes necessary to improve the system. The first method starts and ends with frequency-response data; the second one starts and ends with transfer functions in analytic form. The advantage of knowing the transfer functions is that transient response can be calculated directly.

In the frequency-response method, the transfer function around the loop is calculated or measured for values of $s = j\omega$ only. The result is plotted either as a polar curve or as separate curves of attenuation and phase shift versus frequency. From these curves, a lead circuit is selected which achieves maximum phase lead at the resonant frequency of the system. The gain selected is such that the maximum amplification of any sinusoidal signal is some reasonable value, such as 1.4.

In the root-locus method, the zeros and poles of the transfer function are plotted on the complex s plane; the locus of values of s is then sketched, making the phase shift of the loop transfer function 180°. Any point on this locus is a root of the characteristic equation if the loop gain makes the transfer function around the loop for that signal have a magnitude of unity. A point is then selected where the damping ratio of the dominant complex roots is some reasonable value, such as 0.5. A direct way of selecting a lead circuit is to note that 90° phase lead can be obtained along a damping-ratio line of 0.6; then the locus for the remainder of the system is plotted for a phase shift of 270°. The intersection of this locus with the line for damping ratio of 0.6 is then the location of the root, which determines the time constants of the lead circuit and the loop gain.

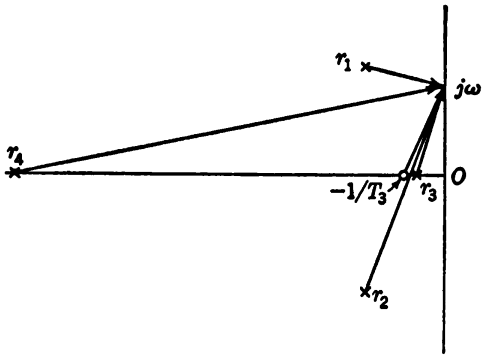


FIG. 7.27. Obtaining the frequency response of closed-loop system using vectors from zeros and poles of closed-loop function.

Note that the fundamental problem in the two methods is the same: to locate the lead circuit that has maximum effectiveness for the signal with which one is finally most concerned. Some trial and error would be involved in either method in the normal procedure. This disadvantage is not necessarily serious since, for common systems, one can soon judge the right value quickly and the choice is not critical. A direct method can be developed with either approach if the criterion in the frequency-response method is selected to be a definite phase margin and if the criterion in the root-locus method is selected to be a given damping ratio for a pair of roots.

The result in the first method is a curve of frequency response of the closed-loop system; the result in the second method is a plot of the zeros and poles of the closed-loop transfer function. These results may be correlated; the direct way is from the zeros and poles to the frequency-response plot. Thus consider the zero and poles of the closed-loop transfer function and the vectors to a point on the $j\omega$ axis shown in Fig. 7.27.

The poles are labeled as r_1, r_2, r_3, and r_4 since they are the roots that were determined in Fig. 7.26. The only zero is that from the lead network, $-1/T_3$. The frequency response of the closed-loop system is found by taking the product of vectors from each of the zeros and poles to a point along the $j\omega$ axis. The maximum amplification will occur for a signal whose frequency is slightly less than the natural frequency of the root r_1.

The transient response can be plotted directly from the zeros and poles, since a transient of the form Ae^{st} exists for each root and the amplitude A can be computed from the plot, using the techniques discussed in Art. 7.9. If the frequency response alone were known, however, the transient response could be found by some form of evaluation of the Fourier integral.

CHAPTER 8

MULTIPLE-LOOP SYSTEMS

A multiple-loop system involves feedback of several signals rather than just one signal. Many systems are stabilized, for example, by feeding back a tachometer signal proportional to the speed of the output shaft as well as the position of the shaft. The procedure in solving such a system by the root-locus method involves solving the loops in succession. A closer examination of this procedure indicates that a block diagram can be set up in a form which permits a parameter, such as the amplifier time constant, to be used as the variable for a root-locus plot. For more involved systems, such as the vibration absorber, the loops are interlinked, but some manipulation of the block diagram will uncouple the loops, permitting a solution by a series of plots. For a system which involves complete intercoupling of three variables, methods using determinants are preferable, as discussed in Chap. 10.

8.1. Tachometer Feedback Loop

A system with tachometer feedback can be solved by two root-locus plots in sequence. A choice in the parameters to be used as the variable in the plots is permissible.

Consider the system as shown in Fig. 8.1. The root-locus plot of Fig. 7.8 determined the poles of the transfer function from the signal θ_a to the output θ_o for various values of K. This permits the inner loop of Fig. 8.1 to be replaced by a single block in factored form, as shown in Fig. 7.18, in which θ_i corresponds to θ_a in Fig. 8.1. The roots of the characteristic equation of the complete system are now the values of s

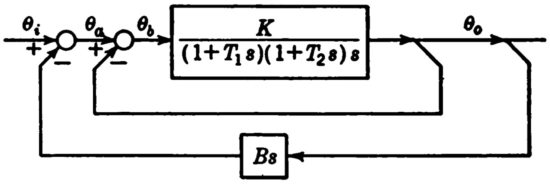


FIG. 8.1. Basic system with tachometer feedback added.

which make the transfer function around the outer loop equal to −1.

$$\frac{Bs}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)} = -1 \tag{8.1}$$

The locus of roots is now plotted with the tachometer constant B variable, as shown in Fig. 8.2, based upon letting $K = K_2$ in Fig. 7.8. Note that for $B = 0$, the roots of the system are at the poles of the forward function, r_1, r_2, and r_3. For a very low value of B, the roots must be somewhere near these poles. In searching for a point s on a small circle around the pole at r_1, the phase angles of the vectors from the poles at r_2 and r_3 and from the zero at the origin are all essentially constant. Therefore, the position of s is selected so that the angle of the vector $s - r_1$ makes the sum of the phase angles of the transfer function around the loop equal 180°. An aid is to consider the effect of these angles in succession; starting with the pole at r_1 and considering it the only pole in the transfer function, the point s would have to be on the horizontal line to the left of r_1 in order that the angle of the vector $s - r_1$ be 180°. The zero at the origin and the pole at r_2 contribute a net lead angle of ϕ_1, however; therefore, the $s - r_1$ vector must be rotated counter clockwise through the angle ϕ_1, to provide an equal and opposite lag angle. Finally, the vector from the pole at r_3 contributes a lag angle of ϕ_3; therefore, the vector $s - r_1$ is rotated clockwise through ϕ_3. As the value of the tachometer constant B increases, the roots move farther out on the locus. Greater lag angles are contributed by the poles at r_2 and r_3, making the locus curve to the right to reduce the lag angle of the vector $s - r_1$, and the locus approaches a vertical asymptote.

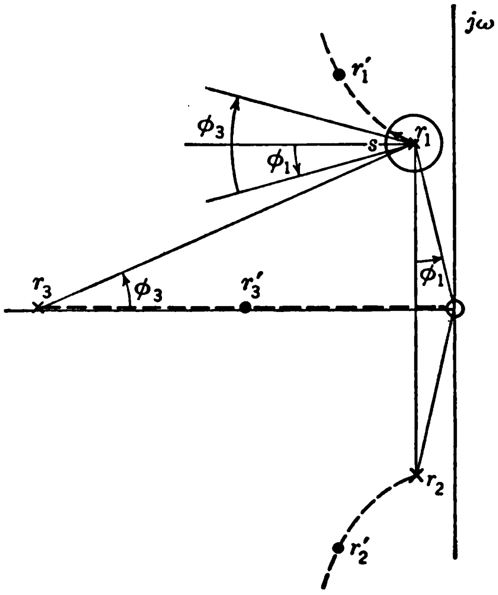


FIG. 8.2. Root-locus plot for system in Fig. 8.1 with tachometer gain variable.

Suppose that the tachometer constant B is selected first. The effect of the gain K can then be studied by adding the two feedback signals to form a single feedback signal $1 + Bs$. The locus of roots versus the gain K must now satisfy the equation

$$\frac{K(1 + Bs)}{(1 + T_1 s)(1 + T_2 s)s} = -1 \tag{8.2}$$

If the gain K is selected to be exactly the same value as in Fig. 8.2, then the roots of the system r'_1, r'_2, and r'_3 must be exactly the same as in Fig. 8.2 for the same value of B. The sequence of plots is not significant.

8.2. Effect of Amplifier Time Constant

> *A block diagram can be set up with any parameter isolated in a separate path. The effect of that parameter can thus be studied by making it the variable for a root-locus plot.*

The basic cubic system shown in Fig. 5.1 often has the motor time constant set by the power requirements and the loop gain K set by error specifications. In selecting the amplifier, however, some choice may exist as to the type of amplifier. The question then arises: What time delay of the amplifier can be tolerated without the performance of the servo being appreciably affected?

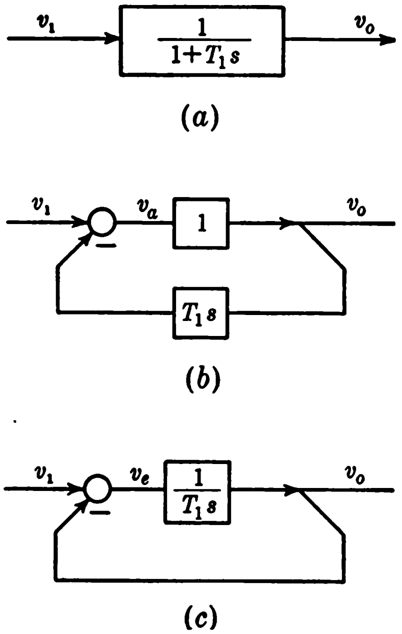


FIG. 8.3. Alternate forms of representation of the transfer function $1/(1 + Ts)$: (a) single block; (b) velocity feedback; (c) forward integration.

In the previous article, the effect of the tachometer constant could be studied because this parameter appeared in the form of a loop gain. Maintaining this concept, the first step is to isolate the quantity T_1 in its own path. This can be done by replacing the single transfer function $1/(1 + T_1s)$ by either of the closed loops shown in Fig. 8.3.

The alternate forms can be justified by considering the inverse transfer function between input and output:

$$\frac{V_i}{V_o}(s) = \frac{V_a}{V_o}(s) + \frac{V_f}{V_o}(s) = 1 + T_1s \tag{8.3}$$

$$\frac{V_i}{V_o}(s) = \frac{V_e}{V_o}(s) + 1 = T_1s + 1 \tag{8.4}$$

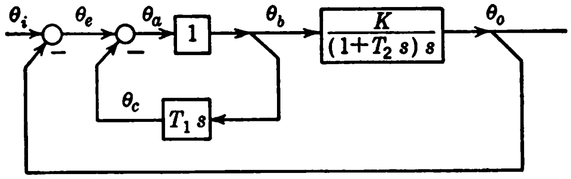
Selecting the diagram shown in Fig. 8.3*b*, the complete block diagram for a cubic system would be as shown in Fig. 8.4.

The effect of T_1 upon the rest of the system is desired; therefore, the order of the loops suggested by the above diagram should be reversed, making T_1s be in the second loop. The inverse transfer function between θ_c and θ_b is

$$\frac{\Theta_c}{\Theta_b}(s) = \frac{\Theta_e}{\Theta_b}(s) - \frac{\Theta_a}{\Theta_b}(s) = \frac{-K}{(1 + T_2s)s} - 1$$

$$= \frac{-(1 - s/r_1)(1 - s/r_2)K}{(1 + T_2s)s} \tag{8.5}$$

Digitized by Google Original from UNIVERSITY OF MICHIGAN



1 + T [illegible]

FIG. 8.4. Modification of Fig. 5.1 to isolate T_1.

The roots of the complete system are now the values of s which make the transfer function around the remaining loop equal to +1.

$$\left[\frac{\Theta_b}{\Theta_c}(s)\right]\left[\frac{\Theta_c}{\Theta_b}(s)\right] = \frac{(1 + T_2 s)s}{-(1 - s/r_1)(1 - s/r_2)K} T_1 s = +1 \qquad (8.6)$$

Note that T_1 is now in the position of a loop gain; the corresponding locus-of-roots plot with T_1 as the variable is shown in Fig. 8.5.

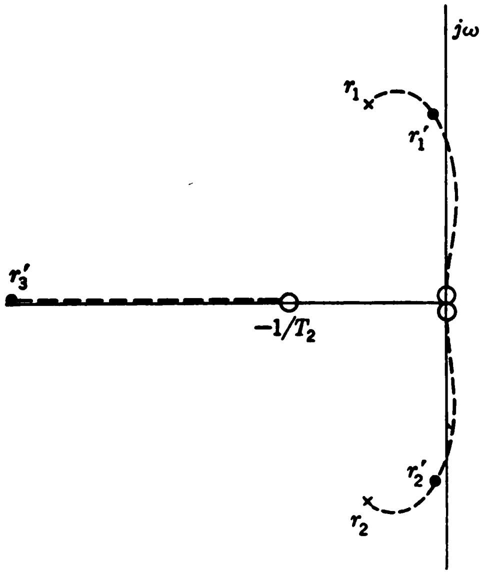


FIG. 8.5. Root-locus plot of the system suown in Fig. 8.4 versus T_1 as the variable.

The excess of zeros over poles in this problem is unusual, but no particular significance should be attached to it. The third root starts from the end of the negative real axis. The main effect of the amplifier time constant is to decrease the damping of the complex roots. The new positions of the roots r_1' and r_2' are for the case when the amplifier time constant is equal to the motor time constant. The distance moved along the locus between r_1 and r_1' is approximately proportional to the time constant T_1 because all vectors in (8.6) except $1 - s/r_1$ are essentially constant. This distance is roughly horizontal to the right, meaning that the decrease in damping constant is approximately proportional to the time constant of the amplifier.

8.3. Effect of Gain Variations from a Fixed Value

Gain variations from a fixed value will occur in a system because of a change in operating conditions. The roots for the fixed value of gain serve as base points for a locus plot versus the variation in gain.

A fixed value of gain is selected on the basis of desired response for a particular set of operating conditions. Variations in gain, however, can occur with a change in operating conditions. Many components, for example, have a linear characteristic within the rated range of signals but saturate for larger signals. An actuator which is balancing a large load torque, for example, will have a small increase in torque for a given increase in signal. Study of the effect of gain variations suggests that gain be separated into two parts, a fixed value K' and a variation ΔK from that value. The general form of the equation being solved is

$$\frac{K(s - q_1)}{(s - p_1)(s - p_2)(s - p_3)} = -1 \tag{8.7}$$

Treating K as $K' + \Delta K$, Eq. (8.7) is more conveniently studied in the form

$$(s - p_1)(s - p_2)(s - p_3) + (s - q_1)(K' + \Delta K) = 0 \tag{8.8}$$

The roots of this equation for $\Delta K = 0$ are known. Denoting these roots by r_i, (8.8) can be rewritten as

$$(s - r_1)(s - r_2)(s - r_3) + (s - q_1)\,\Delta K = 0 \tag{8.9}$$

The sum of the two terms in (8.9) can be made zero by making the ratio of the two terms equal to -1.

$$\frac{\Delta K(s - q_1)}{(s - r_1)(s - r_2)(s - r_3)} = -1 \tag{8.10}$$

Comparison of (8.7) and (8.10) shows that they have exactly the same form; the roots of the equation for $\Delta K = 0$ are the poles of the function for studying gain variations.

Consider the standard cubic system as an example, with the fixed value of gain being that which puts one set of roots on the $j\omega$ axis, as shown in Fig. 8.6. To make the total angle equal 180°, the locus at r_1 must make an angle ϕ_3' with the vertical equal to the angle ϕ_3' of the vector $r_1 - r_3$. For positive values of ΔK, the locus moves up into the right half plane; for negative values, the locus moves down into the left half plane. An extension of this locus for $\Delta K = -K'$ must result in the reestablishment of the original poles of the open-loop transfer function.

Note that the direction of the tangent to the locus curve established at r_1 is a convenient aid in sketching a complete locus plot. The amount of the shift in the roots for $\Delta K/K' = 0.1$ can be estimated by considering the function in (8.10) divided by that in (8.7).

$$\frac{\Delta K}{K'}\,\frac{s - p_1}{s - r_1}\,\frac{s - p_2}{s - r_2}\,\frac{s - p_3}{s - r_3} = 1 \tag{8.11}$$

The value of s in (8.11) is nearly r_1; therefore the factors in (8.11) can be estimated as

Digitized by Google Original from UNIVERSITY OF MICHIGAN

$$\frac{\Delta K}{K'}\frac{r_1}{\Delta s}\frac{1}{2}1 = 1 \tag{8.12}$$

The shift along the locus is thus $\Delta s = \frac{1}{2}r_1(\Delta K/K')$.

The algebraic manipulations involved in this analysis suggest more facts which are useful in making plots, as discussed in the next article.

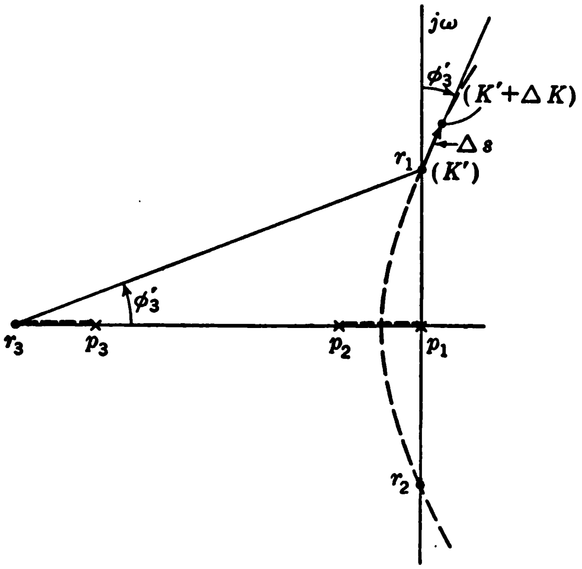


FIG. 8.6. Effect of incremental increase in gain K.

8.4. The Fixed Vector Sum of Roots

The vector ~~sum~~ of the roots is a point along the real axis from which the sum of vectors to all roots is zero. For a system whose loop transfer function has at least two more poles than zeros, the vector sum of the roots is fixed for any and all values of loop gain.

The analysis in the previous article demonstrated that a locus could be continued unchanged if the original poles are replaced by the roots for a given gain. The proof involves (8.8) and (8.9) in which p_i are the poles of the loop function and r_i are the roots for $K = K'$. These equations are identical. Expansion of each of these equations into polynomial form gives

$$s^3 - s^2(p_1 + p_2 + p_3) + s(p_1p_2 + p_1p_3 + p_2p_3 + K' + \Delta K) - p_1p_2p_3 - q_1(K' + \Delta K) = 0 \tag{8.13}$$

$$s^3 - s^2(r_1 + r_2 + r_3) + s(r_1r_2 + r_1r_3 + r_2r_3 + \Delta K) - r_1r_2r_3 - q_1\Delta K = 0 \tag{8.14}$$

Equations (8.13) and (8.14) are also identical; therefore, the coefficients of like powers of s must be the same. In particular, comparing the s^2

terms, the coefficient in (8.13) is the sum of the vectors from the origin to the poles, and the coefficient in (8.14) is the sum of the vectors from the origin to the roots. The equality of these two sums can be expressed by stating that the vector sum of the two sets is the same. This vector sum remains fixed regardless of the value of gain; the only restriction is that the original function must contain at least two more poles than zeros in order that the gain K not be involved in this term.

The fixed position of the vector sum has several uses in many plots. In Fig. 8.6, the distance $r_3 - p_3$ by which the root shifts to the left must equal the distance p_2 by which the root r_2 shifts to the right. This fact establishes r_3 accurately and hence the angle of the locus, ϕ'_3, at the point r_1.

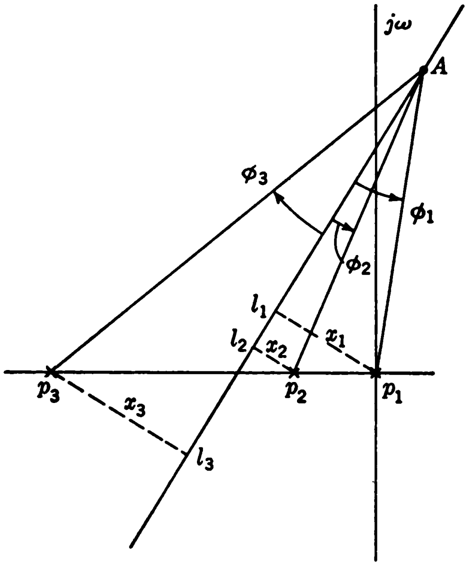


FIG. 8.7. The asymptote through the vector sum of the roots.

In Art. 7.3, the angles of the asymptotes for very large values of s are established on the basis that the angles of all vectors are approximately equal. The asymptotes are of greater value in establishing the locus if they are drawn through the vector sum of the roots, as shown in Fig. 8.7.[1]

The asymptote through the vector sum of the roots lies between the vectors to the point A, so that angles from poles to the right of the asymptote are greater than 60° and the other angles are less than 60°. The deviation of an angle from 60° can be approximated in radian form as the length of the perpendicular to the asymptote divided by the base length measured along the asymptote from A to the perpendicular. When the asymptote passes through the vector sum of the roots, the sum of the perpendiculars is zero; therefore, the sum of the angular deviations would be zero if all the base lengths were the same. The locus therefore approaches the asymptote at a distance sufficient that all the base lengths are approximately equal.

If the loop transfer function contains any zeros, these points should be included in locating the vector sum of the roots for the purpose of drawing asymptotes. The length to a zero is treated as a negative quantity compared to the length to a pole at the same position.

---

[1] This choice for the center of the asymptotes was developed by R. M. Stewart at the California Institute of Technology.

8.5. Classic Problem of Finding Roots of *n*th-degree Equation

The zeros of the sum of two terms are the values of s which make the ratio of the terms equal to −1. An nth-degree polynomial can be factored by a sequence of not more than n/2 root-locus plots.

A linear system is described as a set of differential equations; the block diagram is a form of presentation of these equations which is convenient for control systems. The classic procedure is to combine the equations into a single equation of the form

$$s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = 0 \tag{8.15}$$

The coefficients will, in general, be a complex combination of the original parameters of the system, as demonstrated in Art. 7.1. For such systems a single solution of the equation is of little value in determining the effect of these parameters. Some systems, however, are very complicated and have only a few parameters which can be varied. If these parameters are identified literally and numerical values are substituted for fixed parameters, an equation of the form of (8.15) is obtained in which the coefficients of the terms are easily recognizable combinations of the variable parameters. In one unusual application of a servo, the system was described by a fourth-degree equation in which each coefficient was a loop gain completely independent of the other coefficients.

The zeros of the sum of two terms can be found as the values of s which make the ratio of the two terms equal to −1. This procedure permits the roots of an nth-degree equation to be determined by a succession of plots in which two terms are combined for each plot. The procedure will be illustrated for a simple fourth-degree equation:

$$s^4 + as^3 + bs^2 + cs + d = \{[(s + a)s + b]s + c\}s + d = 0 \tag{8.16}$$

The right side of (8.16) can be shown to be equal to the left side by expansion. The zeros of the function in brackets can be found by the familiar quadratic plot shown in Fig. 8.8*a*.

The quadratic can now be expressed in factored form, making (8.16) become

$$[(s - p_1)(s - p_2)s + c]s + d = 0 \tag{8.17}$$

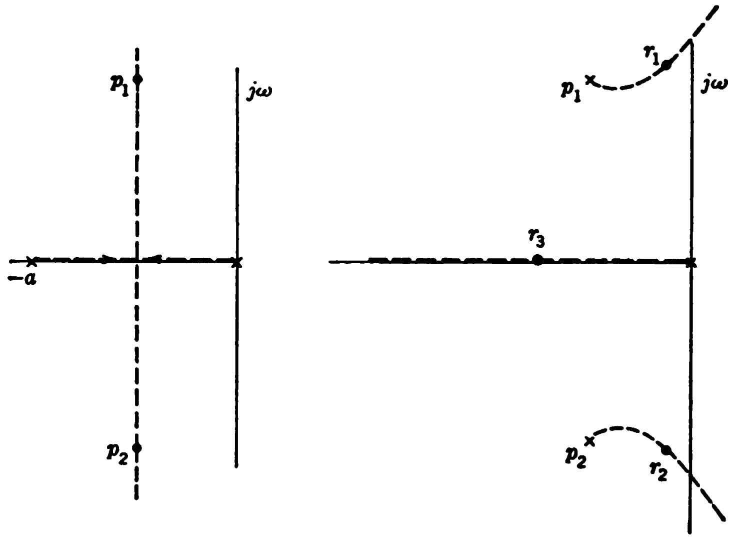
The zeros of the cubic in brackets can then be found in a plot which starts from the end points of the previous plot, as shown in Fig. 8.8*b*. The process could be repeated for a higher-order equation.

This procedure makes each coefficient of the equation the variable in a root-locus plot. The sequence of consideration of the coefficients can be reversed by substituting $s = 1/x$.

The number of plots can be approximately halved by treating two terms at a time, as shown in

$$s^6 + as^5 + bs^4 + cs^3 + ds^2 + es + f$$
$$= \{[(s + a)s^2 + b(s + c/b)]s^2 + d(s + e/d)\}s + f \quad (8.18)$$

One plot determines the zeros of the quantity in the brackets; the second plot finds the zeros of the quantity in braces; and the third plot determines the zeros of the entire function. The total number of plots is thus never more than half of the degree of the equation.



(a) (b)

FIG. 8.8. Finding the zeros of a polynomial by successive plots: (a) neglecting all coefficients except a and b; (b) neglecting all coefficients except a, b, and c.

A particular coefficient can be made the variable in the final root-locus plot by adding its term after the remainder of the function is in factored form. The technique of obtaining a single equation and adding terms one at a time arises in the expansion of determinants in Chap. 10. The block-diagram technique, however, is particularly well suited to isolating the effects of physical parameters of a system and can be applied to any type of system, as illustrated in the next article.

8.6. Vibration Absorber

A vibration absorber is a spring-mass system which is mounted on another mass to absorb the vibration at a particular frequency. The analysis can be carried out in block-diagram form and suggests a means of stabilizing a system having a resonant component.

A mechanical diagram of a vibration absorber is shown in Fig. 8.9. The large mass M_1 is typically a unit, such as a turbine, which is subject to a driving force of constant frequency due to unbalance of the rotating member. The vibration absorber is the small spring-mass combination attached to the turbine. The natural frequency of the vibration absorber itself as a single unit is made equal to the driving frequency ω_d, with the result that the mass M_2 transmits a force to the large mass M_1 equal and opposite to the applied force.

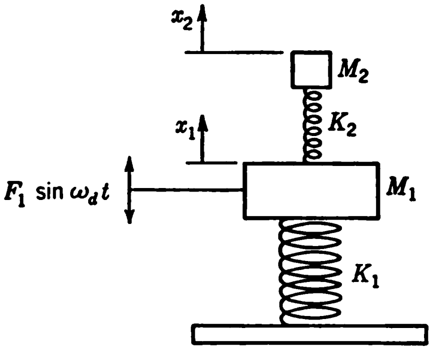


FIG. 8.9. The vibration-absorber system.

The block diagram of such a system (Fig. 8.10) can be set up step by step as follows: The main line of action of the signals is set up first, with the feedback point of undetermined quantities merely indicated. Thus the applied force $F_1 \sin \omega_d t$ minus the coupling force f_{12} from the spring K_2 and the restoring force from the spring K_1 is the net force acting on the mass M_1. This net force is equal to the mass times acceleration of the mass M_1. The displacement x_1 is obtained by multiplying the net force by the transfer function $1/M_1s^2$. Subtracting the motion x_2 from x_1 leaves the stretch of the small spring, which, when multiplied by the spring constant K_2, gives the force acting between the two masses, f_{12}. This force is the only external force acting on M_2; therefore, multiplying by the transfer function $1/M_2s^2$ gives the displacement x_2.

Now all the quantities are present in the block diagram (see Fig. 8.10), from which the previously indicated feedback signals may be obtained. If some signal has been overlooked in the process, it can be included very easily. Thus if some damping is present in the mount for M_1, then an

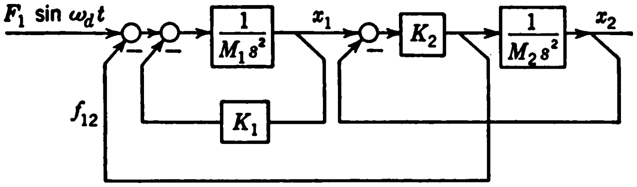


FIG. 8.10. Block diagram of vibration absorber.

extra signal Dsx_1 would be fed back from the displacement x_1 to the input point.

The block diagram can be simplified by thinking in terms of superposition of signals. The total force acting on M_1 due to the motion x_1 is the sum of the forces due to K_1 and K_2. The force on M_1 due to x_2 is K_2x_2. A block diagram suggests itself with two simple spring-mass systems having a force on M_1 due to motion x_2, as shown in Fig. 8.11.

The inner loops are simple resonant systems for which the transfer functions have poles along the $j\omega$ axis. Closing the outer loop introduces

Digitized by Google Original from UNIVERSITY OF MICHIGAN

the effect of the force acting on the mass M_1 due to the motion x_2. The root-locus plot with K_2 as the loop gain is shown in Fig. 8.12. The low natural frequency is driven lower, and the high natural frequency is driven higher. This result can be checked by physical reasoning directly from the mechanical diagram of Fig. 8.9 by noting that for the low-frequency vibration of M_1, the motion of M_2 is in phase and makes K_2 appear to be a weaker spring than if M_2 were fixed.

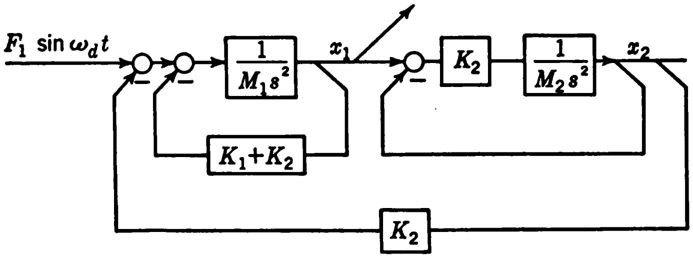


Fig. 8.11. Modification of Fig. 8.10 to isolate individual spring-mass systems.

The transfer function from the input, denoted in Fig. 8.10 as $F_1 \sin \omega_d t$, to x_1 is given by

$$\frac{X_1}{F_1}(s) = \frac{(1 + M_2 s^2/K_2)/K_1}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)(1 - s/r_4)} \tag{8.19}$$

The zeros of this function are $\pm j\sqrt{K_2/M_2}$, because these points are poles of the feedback function in Fig. 8.11. Suppose that the driving frequency ω_d is not fixed but may have a small deviation from its normal value. The frequency response of the system can then be obtained by the product of vector lengths to the point $s = j\omega_d$.

The displacement x_1 is zero if the driving frequency is exactly equal to the given frequency $\sqrt{K_2/M_2}$ of the vibration absorber. For slight deviations of the driving frequency ω_d from this value, the amplitude is proportional to the deviation. If the driving frequency happens to be equal to the resonant frequency r_1 of the system, the oscillation would build up to amplitudes of oscillation limited only by the nonlinear behavior of the mounts. If some variation of the driving frequency is expected, therefore, the spring K_2 will have to be large enough

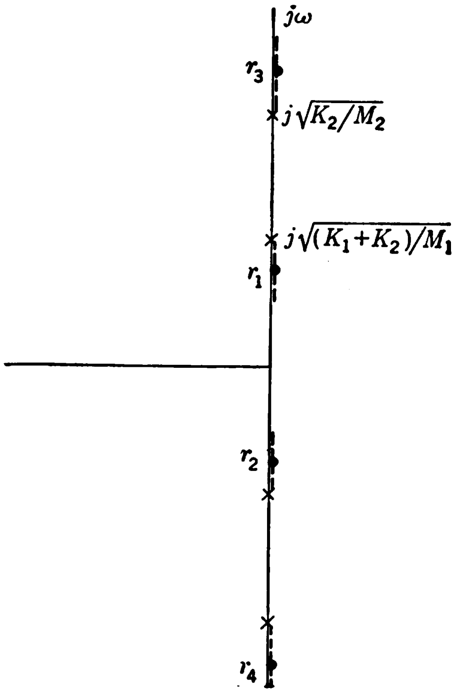


Fig. 8.12. Root-locus plot for system in Fig. 8.11 versus K_2 in the outer loop.

that the root r_3 is shifted by a safe distance from the zero at $j\sqrt{K_2/M_2}$. The vibration damper demonstrates a principle which will prove useful in control systems which have a similar characteristic.

8.7. Stabilization of System with Resonant Inner Loop

> *A system with a resonant inner loop can be stabilized by addition of phase lags so that* 180° *phase shift is introduced in the outer loop at the resonant frequency of the inner loop. A bridge circuit is preferable to multiple-lag circuits because less attenuation of the damping signal is involved.*

Consider the system shown in Fig. 8.13. The input is the gauge pressure P in the bellows, and the output is the shaft displacement x_o of the free end of the spring. At equilibrium, the force on the mass M due to the gauge pressure in the bellows is balanced by the spring force. With the servo loop open, this spring-mass system has a natural resonant frequency. With the servo loop closed, a pick-off signal x_a acts through the transfer function $KG(s)$ of the amplifier and motor to change the output angle, which in turn produces a force on the mass through the spring. A block diagram may thus be constructed with the inner loop resonant and the servo in the outer loop, as shown in Fig. 8.14.

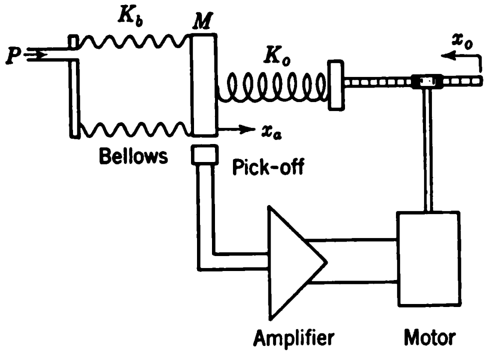


FIG. 8.13. A system with a resonant component.

If the transfer function $G(s)$ is simply $1/s$, then the locus of roots of the outer loop versus gain K is as shown in Fig. 8.15. The root r_1 moves directly to the right from the pole at $j\omega_0$, because 180° phase shift is already present at that point, arising from vectors from the poles at the origin and at $-j\omega_0$. This initial direction of the locus could be reversed if an additional 180° phase shift were introduced. A simple way to achieve this phase shift is to make $G(s) = 1/(1 + Ts)^3s$, in which $T = 2/\omega_0$ (Fig. 8.16).

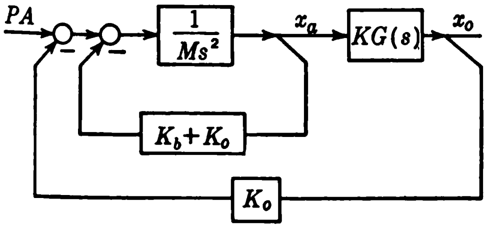


FIG. 8.14. Block diagram for system in Fig. 8.13.

The locus from $j\omega_0$ is now reversed, and the gain may be increased until the new roots which break away from the real axis cross the $j\omega$ axis. Consider, however, the amount of damping obtained in the root r_1 if

Digitized by Google Original from UNIVERSITY OF MICHIGAN

the gain is limited to critical damping on the roots r_3 and r_4. Each term $1/(1 + Ts)$ introduces an attenuation of about 2 for $s = r_1$ but an amplification of about 1.3 for $s = r_3$. The triple delay $1/(1 + Ts)^3$ then introduces an attenuation of 8 at $s = r_1$ and an amplification of 2.2 at $s = r_3$. From this cause alone the length of the vector $r_1 - j\omega_0$ is limited to one-twentieth the length of the vector from the origin to r_3. To achieve greater damping in r_1, select a function $G(s)$ which gives phase shift without attenuation of signal. This result is achieved by use of a phase-shifting bridge circuit for which the transfer function from input v_i to output v_o is given by

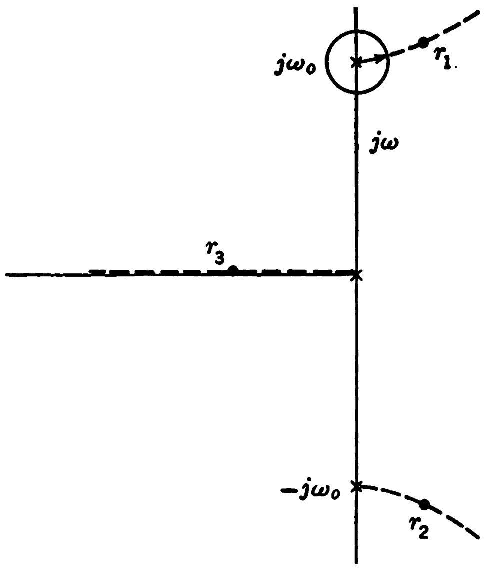


FIG. 8.15. Root-locus plot for system in Fig. 8.14 with $K = K_o$ as loop gain and $G(s) = 1/s$.

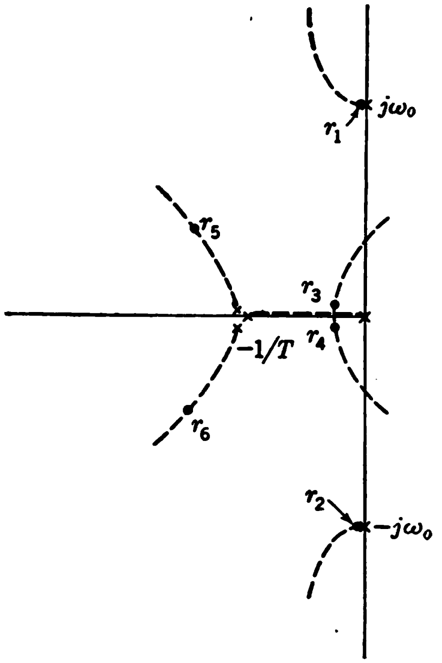


FIG. 8.16. Root-locus plot for system in Fig. 8.14 with $G(s) = 1/(1 + Ts)^3s$.

$$\frac{V_o}{V_i}(s) = \frac{1 - Ts}{1 + Ts} = -\frac{(s - 1/T)}{(s + 1/T)} \qquad T = RC \tag{8.20}$$

Selecting T for 120° phase shift at $j\omega_0$, the corresponding locus of roots is as shown in Fig. 8.17.

Greater damping can, of course, be obtained by adding one simple time delay to bring the total phase angle to 180°. The system in which this circuit has been used, however, has additional delays in the system not specifically shown in the block diagram, which bring the total phase shift to 180°.

If the value T is to be selected for maximum gain before any root crosses the $j\omega$ axis, the roots from $\pm j\omega_0$ and the roots from the real axis meet at points in the left half plane. Still higher gain would cause one set of roots to cross the $j\omega$ axis. The gain is maximum for this system because the attenuation from the frequency-dependent portion of the transfer function is the maximum possible for any point between 0 and $j\omega_0$.

The plot near the point of intersection of the loci is not critical; any point in this region will have a measured phase-shift angle of nearly 180°. The expression for transient response involves a special case, which is considered in Chap. 11. Further improvement of the system requires the use of an active circuit.

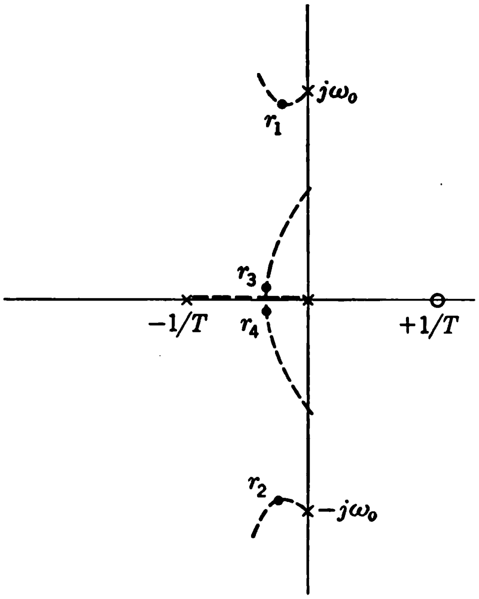


FIG. 8.17. Root-locus plot for system in Fig. 8.14 with $G(s) = (1 - Ts)/(1 + Ts)s$.

8.8. The Active-compensation Technique

> *"Active compensation" involves an amplifier in the forward path and a passive circuit in the feedback path of a loop. For high loop gain, the over-all transfer function is approximately equal to the reciprocal of the feedback function.*

Consider the system shown in Fig. 8.18. The transfer function from the output of the amplifier v_o to the feedback voltage v_f is given by

$$\frac{V_f}{V_o}(s) = \frac{1/Cs}{1/Cs + R + Ls} = \frac{1}{1 + RCs + LCs^2}$$
$$= \frac{1}{(1 - s/q_1)(1 - s/q_2)} \qquad (8.21)$$

The input resistors R_1 are equal; therefore, the input to the amplifier is the average of the input voltage v_i and the feedback voltage v_f. Negative feedback is achieved by having the amplifier have an odd number of stages so that the output is of opposite sign from the input.

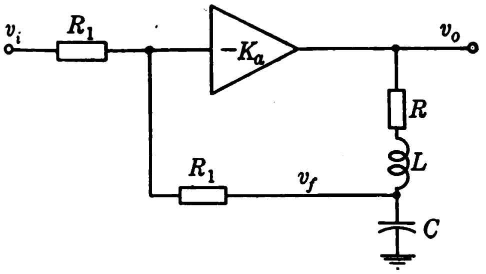


FIG. 8.18. An active-compensation network.

The root-locus plot for the loop satisfies the condition

$$\frac{K_a}{(1 - s/q_1)(1 - s/q_2)} = -1 \qquad (8.22)$$

The roots having exactly the same

damping as the poles, the natural frequency increases with loop gain K_a.

The transfer function from input to output for this closed loop is then

$$\frac{V_o}{V_i}(s) = -\frac{(1 - s/q_1)(1 - s/q_2)}{(1 - s/r_1)(1 - s/r_2)}\,\frac{K_a}{K_a + 1} \tag{8.23}$$

Consider now the use of this active network as $G(s)$ in the system shown in Fig. 8.14. As a first trial, select the natural frequency of the zeros q_1 and q_2 to be equal to that of the poles $\pm j\omega_0$. The corresponding locus-of-roots plot is shown in Fig. 8.19.

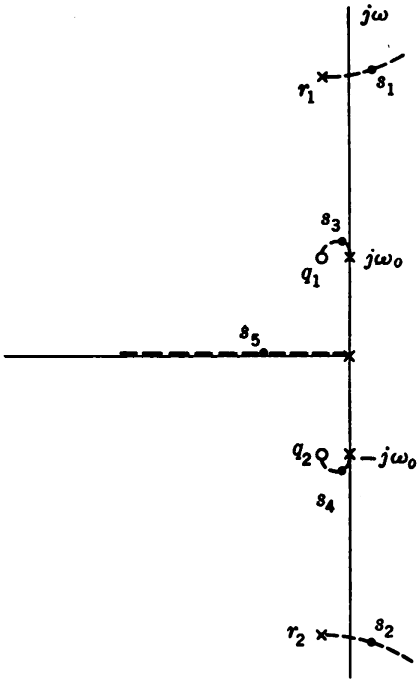


FIG. 8.19. Root-locus plot for system in Fig. 8.14 with $G(s)$ given by the active-compensation network in Fig. 8.18.

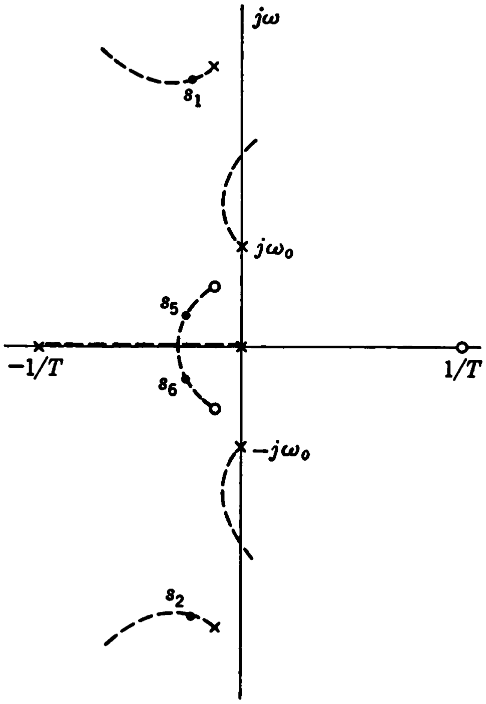


FIG. 8.20. Root-locus plot for the system in Fig. 8.14 with $G(s)$ given by the active-compensation network of Fig. 8.18 and $(1 - Ts)/(1 + Ts)$.

The locus starts vertically from $j\omega_0$ and curves around in a semicircle to q_1. The locus from r_2, on the other hand, heads to the right like the locus from $j\omega_0$ in the uncompensated system shown in Fig. 8.15. If the loop gain is to be limited by these roots, then the zeros q_1 and q_2 should be selected with a lower natural frequency than the poles at $\pm j\omega_0$ in order that the locus from $\pm j\omega_0$ may start to the left.

Actually the type of correction used in Fig. 8.17 can be used in this case in order to reverse the locus from r_1. This bridge circuit will introduce phase shift in the region of $j\omega_0$ also; thus the natural frequency of the zeros should be slightly less than ω_0. The root-locus plot for such a system is as shown in Fig. 8.20.

Results such as this tend to imply that any undesirable characteristic of a linear system can be corrected by introducing a compensating characteristic in series. In so far as the assumptions of linearity and absence of unwanted signals are justified, this conclusion is true. However, such systems are susceptible to a return to the original characteristic when some condition such as saturation makes the linearity assumption void. Tests on a system whose linear analysis corresponds to the plot of Fig. 8.17, for example, match this behavior quite well for normal amplitude of signals. A very large signal, however, maintains continuous oscillation. Such test results are mentioned in order to avoid giving the impression that any characteristic can be obtained by adding a sufficient number of amplifiers and networks.

These examples do provide a background of block-diagram manipulation and characteristics of corrective networks, however, which are helpful in attacking a more complex problem, such as autopilot design.

8.9. Roll Autopilot Control

An autopilot incorporates all the sensing elements and control-surface actuators necessary to maintain an airplane in straight and level flight in spite of disturbing action such as wind gusts.

The simplest function of an autopilot is to maintain the plane level in roll. A block diagram of the system is shown in Fig. 8.21. A gyro provides a horizontal reference position from which the error angle θ_e of the plane in roll is measured. This signal input to a controller with transfer function $KG(s)$ provides an output which is the desired angle of the control surfaces, λ_i. The servo actuator drives the ailerons to an angle λ_o. The aileron deflection causes a torque on the plane which determines the change in roll angle of the plane, θ_o. In analyzing this system, the transfer functions of each component are assumed to be of the form shown in Fig. 8.21.

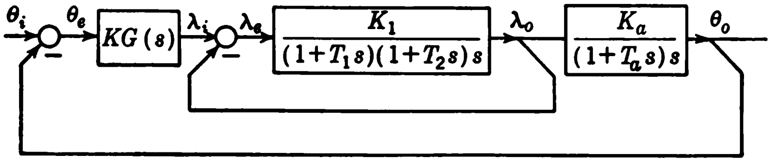


FIG. 8.21. The roll autopilot system.

The servo loop can be replaced by the single block shown in Fig. 7.18. Assuming the values of r_1, r_2, r_3, and $-1/T_a$ shown in Fig. 8.22 with $G(s) = 1$, the locus versus KK_a is as shown by the dashed lines. If a higher gain is needed in this system, a lead network may be added to bend the low-frequency locus away from the j axis. The network with $G(s)$ equal to $(1 + Ts)/(1 + \alpha Ts)$, having the values shown in Fig. 8.23, produces a new plot in which the two loci meet and then break away. The gain is limited for such a system, because the locus is sure to cross the j axis below r_1. Increasing the gain within the servo loop would raise the natural frequency of r_1 but would also decrease the damping of that root. Inserting a lead network within the servo loop, however, would increase the damping of r_1, and the zeros of the lead network would also be present in the transfer function around the outer loop. In selecting this lead network, it is convenient to combine the two feedback paths around the servo function, making the transfer function from λ_o to λ_e become

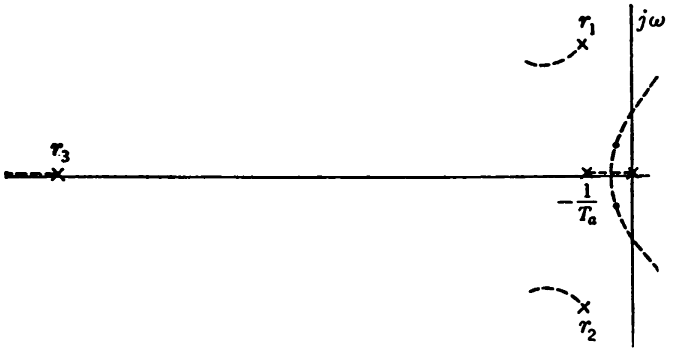


FIG. 8.22. Root-locus plot of system in Fig. 8.21 versus KK_a with $G(s) = 1$.

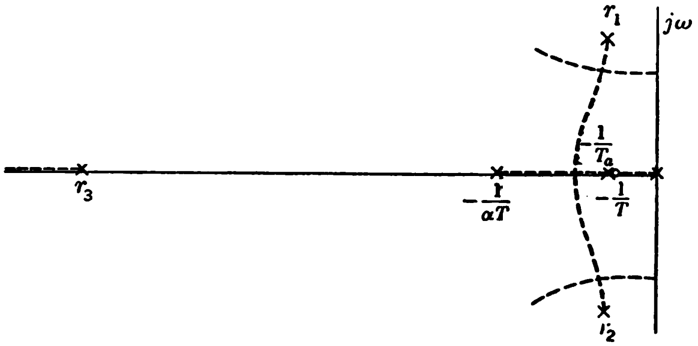


FIG. 8.23. Root-locus plot of system in Fig. 8.21 versus K with $G(s) = (1 + Ts)/(1 + \alpha Ts)$.

$$\frac{\Lambda_e}{\Lambda_o}(s) = 1 + \frac{K_a}{(1 + T_a s)s} = \frac{K_a(1 - s/q_1)(1 - s/q_2)}{(1 + T_a s)s} \tag{8.24}$$

The zeros of this function, q_1 and q_2, are the values of s which make the ratio of the two terms equal to -1, and they can be found by a simple quadratic plot.

Replacing the two feedback paths by their single equivalent, and using the lead network in the servo function, the loop transfer function is then

$$\frac{K_1}{(1 + T_1 s)(1 + T_2 s)s} \frac{(1 + Ts)}{(1 + \alpha Ts)} \frac{K_a(1 - s/q_1)(1 - s/q_2)}{(1 + T_a s)s} = -1 \tag{8.25}$$

The resultant root-locus plot is shown in Fig. 8.24. For frequencies high compared to q_1 the poles and zeros near the origin cancel. The plot then is based upon the pole at $-1/\alpha T$ rather than a pole at the origin as in Fig. 7.8.

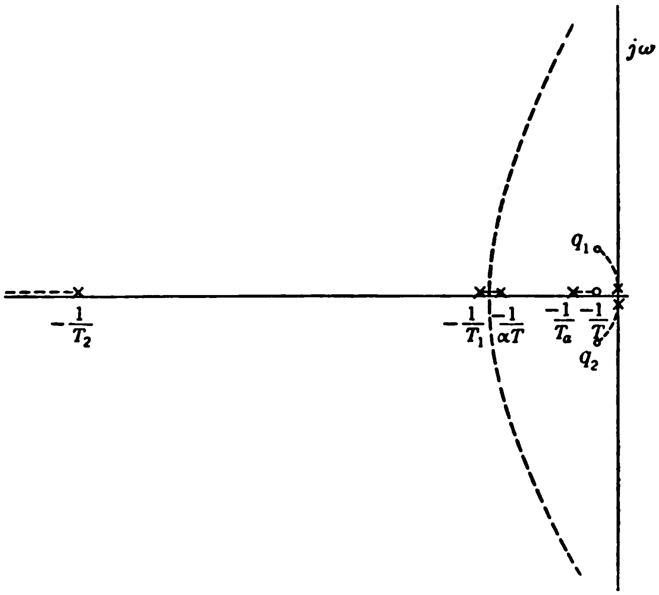


FIG. 8.24. Root-locus plot of system in Fig. 8.21 versus K_1 with $(1 + Ts)/(1 + \alpha Ts)$ in inner loop.

8.10. Pitch Autopilot Control

A pitch autopilot must control altitude as well as attitude of the airplane. The basic block diagram involves almost complete intercoupling of loops; a solution is obtained by considering a locked-control operation first and then combining the two control functions.

The variables involved in a pitch autopilot system are shown in Fig. 8.25. All angles are variations from their steady-state values for straight and level flight. The change in angle of attack, α, is then the change in attitude angle of the plane, γ, minus change in the angle of the velocity vector with respect to horizontal, β. This angle of attack determines the net torque and lift forces which act on the airplane, owing to aerodynamic forces, when the controls are locked. The complete block diagram of the system (Fig. 8.26) can be built up in simple steps.

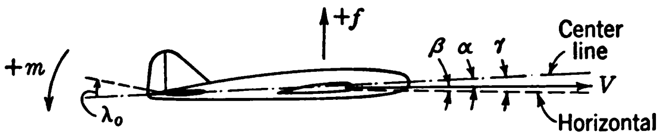


FIG. 8.25. The pitch autopilot system.

Consider first the main line of action of signals from the input signal λ_i to the final output β. The servo moves the elevators through the angle λ_o, which in turn causes a torque on the airplane, m_i. The resultant acceleration, limited by the inertia J, changes the attitude angle of the plane, γ. The angle of attack α also changes, producing a lift force on the plane f. This force produces an acceleration and hence vertical velocity, or it changes the angle β of the total velocity vector.

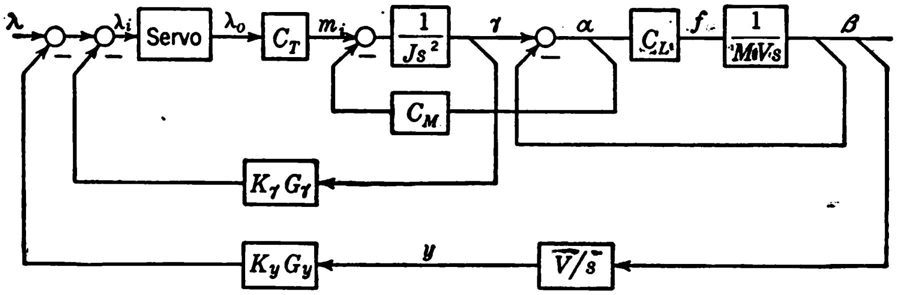


FIG. 8.26. Block diagram of system in Fig. 8.25 written directly from inspection.

The natural feedback signals are the restoring torque $C_M\alpha$ on the plane due to angle of attack α, and the subtraction of β from γ to give α. The servo signals come from a gyro measurement of attitude, $K_\gamma G_\gamma \gamma$, and a pressure or radar measurement of altitude, $K_y G_y y$. The complete block diagram involves intercoupling of loops to the extent that no signal loop can be solved directly. The emphasis in such a system thus shifts to the pattern of simplification required to obtain a solution.

If the control loops are opened, $K_y G_y = 0$, $K_\gamma G_\gamma = 0$, corresponding to locking the control surfaces, the remaining block diagram is of somewhat the same form as that of Fig. 8.10. Using superposition of signals in the same manner, the block diagram can be simplified to the form shown in Fig. 8.27.

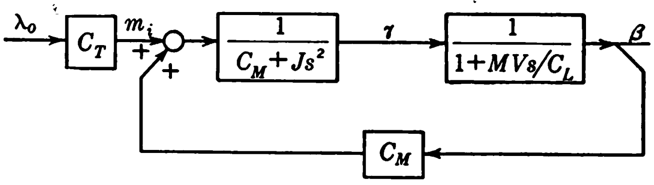


FIG. 8.27. Simplification of the block diagram in Fig. 8.26 for the special case of locked controls.

The loop gain for this system is fixed at unity; the resultant root-locus plot is shown in Fig. 8.28. For the loop gain of 1 involved in this loop, the root from $-C_L/MV$ moves to the origin, and the roots from the $j\omega$ axis move to the left by approximately $C_L/2MV$.

In setting up the complete block diagram again, note that the angle γ is simply $1 + (MVs/C_L)$ times the angle β; therefore, the attitude control signal can be taken from β, as shown in Fig. 8.29.

Note that the control functions are in parallel and that a feedback signal produced by one could have been produced by the other. To minimize the complexity of the system, consider first the total feedback which can be obtained by letting both G_γ and G_y be 1 and varying the ratio of gains K_y/K_γ.

$$K_\gamma\left(1 + \frac{MV}{C_L}s\right) + \frac{K_yV}{s} = \frac{K_yV}{s}\left(1 - \frac{s}{q_1}\right)\left(1 - \frac{s}{q_2}\right) \qquad (8.26)$$

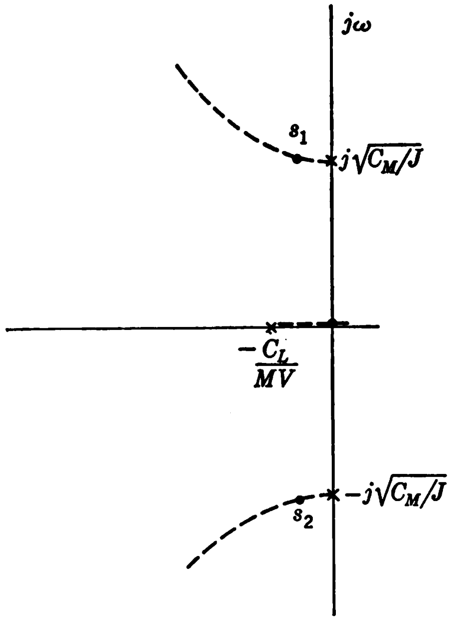


Fig. 8.28. Root-locus plot of the system in Fig. 8.27.

The zeros of this combined feedback function can be found by plotting the locus of values of s which make the ratio of the two individual functions on the left side of (8.26) equal to -1. The locus plot is that of a typical quadratic, and the zeros have a damping constant $-C_L/2MV$, which is equal to that of the poles s_1 and s_2 of the locked-control transfer function. By proper selection of the ratio of gains K_y/K_γ the zeros q_1 and q_2 can be made to nearly coincide with the poles s_1 and s_2. Since $|s_1|^2$ is approximately equal to C_M/J, the required gain ratio K_y/K_γ is $(M/C_L)(C_M/J)$. After setting this gain ratio, the combined feedback signal is then passed through a lead network $K_C(1 + Ts)/(1 + \alpha Ts)$. Denoting the servo transfer function as $(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)$, the loop transfer function for the system in Fig. 8.29 is

$$\frac{1}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)}\,\frac{C_T/C_M}{(1 - s/s_1)(1 - s/s_2)s} \cdot \frac{K_yV(1 - s/q_1)(1 - s/q_2)}{s}\,\frac{K_C(1 + Ts)}{1 + \alpha Ts} = -1 \qquad (8.27)$$

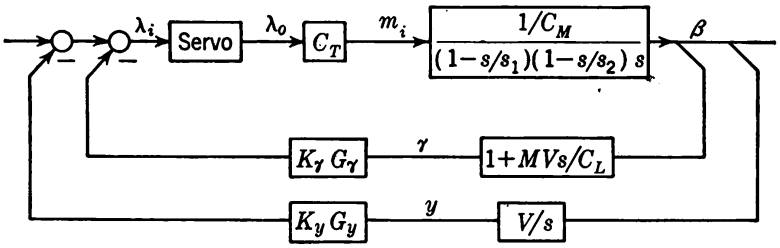


Fig. 8.29. Modification of block diagram of Fig. 8.26 to use locked-control transfer function obtained in Fig. 8.28.

With all values as shown on Fig. 8.30, the locus of roots versus loop gain is shown by the dashed lines. The lead network provides the necessary damping for the roots p_1 and p_2. The pair of roots which move from s_1 and s_2 toward q_1 and q_2 mean that the natural response of the airplane denoted by s_1 and s_2 is not appreciably influenced by the control loop.

This solution is not necessarily the best solution for the system; it merely illustrates the procedure by which the block diagram may be manipulated in order to solve such a system. The solution is not yet complete, in that the roots constitute the poles of the transfer function between any two points but the zeros depend upon the input and output point. The zeros can often be found by recalling that the zeros of the transfer function from an input to an output of a single-loop system are the zeros of the forward function and the poles of the feedback function.

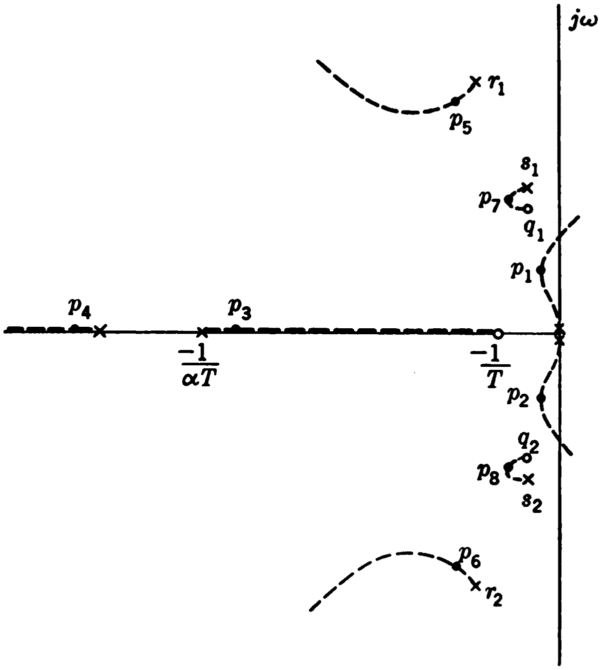


FIG. 8.30. Root-locus plot for system in Fig. 8.29 with combined feedback functions.

8.11. The Problem Presented by Multicoupling

> *A multicoupled system is one which is described by a set of equations in several variables. The block diagram can be manipulated to obtain a solution, but more powerful methods are desirable.*

The block diagram of the pitch autopilot involves intercoupling of loops, but the natural procedure of considering locked-control performance first provides a simplification sufficient to permit a solution. Suppose, however, that the deflection of the elevators produces a force on the airplane which is not negligible. This would require an additional signal path in Fig. 8.26 from λ_o acting through C_R (not shown) to add to $C_L\alpha$. The three variables γ, β, and λ_o are now each involved in three equations: those for torque, lift, and control. Such a system is said to be multicoupled. The definition is not precise, but the term multicoupled system is used in this text to describe a system in which the primary problem is that of establishing the pattern of solution, as distinguished from a multiloop system, in which the sequence of closing loops is reasonably apparent.

The system described above will be reduced to a multiple-loop system, however, and then to a single transfer function between two points.

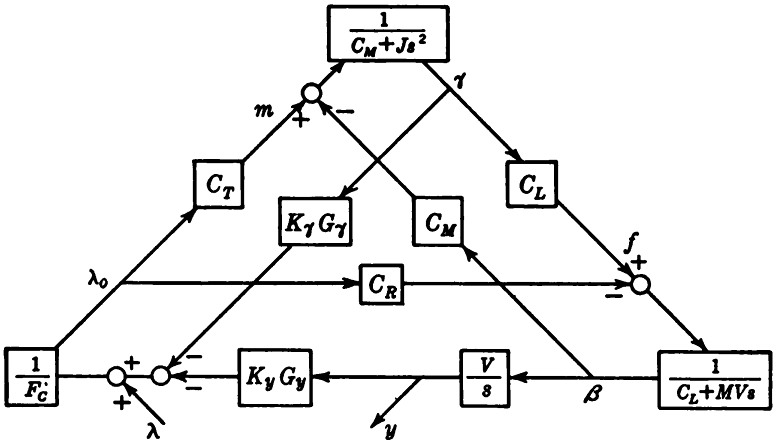


FIG. 8.31. Block diagram of Fig. 8.26 modified to include all possible intercoupling and arranged in symmetrical form.

The block-diagram reasoning may then be correlated with other techniques discussed in Chaps. 9 and 10. The first step is to develop a block diagram in which the symmetry of the flow of signals is emphasized, as shown in Fig. 8.31.

The main line of action of signals is clockwise around the triangle. The desired elevator angle λ acting through the servo transfer function produces the actual elevator angle λ_o. This angle causes a torque on the plane and a resultant attitude angle γ. The angle γ causes a lift force f to be present, which changes the angle of the velocity vector of the plane, β. The resultant vertical displacement is then measured and acts through the control function K_yG_y to give a signal to the servo, thus closing the main loop. The coupling effects present in this system introduce additional signal paths as shown. These signals are not simple feedbacks, in that all the loops are linked. Simplification in form can be obtained by treating only the angles λ_o, γ and β as variables and identifying transfer functions by subscripts, as shown in Fig. 8.32.

This modification involves only the multiplication of the transfer functions between two angles in each case. The variables θ_1, θ_2, and θ_3 correspond to λ_o, γ, and β, respectively. The transfer functions are identified by their positions in the block diagram.

In finding the transfer function from θ_x to θ_3, the variable θ_2 need not be preserved as a distinct quantity. The flow of signals at that point can therefore be clarified by identifying each path as shown in Fig. 8.33. The signal $a_{21}\theta_1$, for example, feeds back on θ_1, as well as going on through a_{32} to help produce θ_3. In particular, the path from θ_3 through a_{23} and a_{12} to θ_1 is simplified by avoiding the crossover pattern in Fig. 8.32. The transfer functions, including feedback loops and combining parallel paths, take the form of a single main loop, as shown in Fig. 8.34. Each function is obtained in factored form by a root-locus plot. The final loop plot involves the complete complexity of the system, but after these roots are found, the transfer function from θ_x to θ_3 may be expressed in factored form.

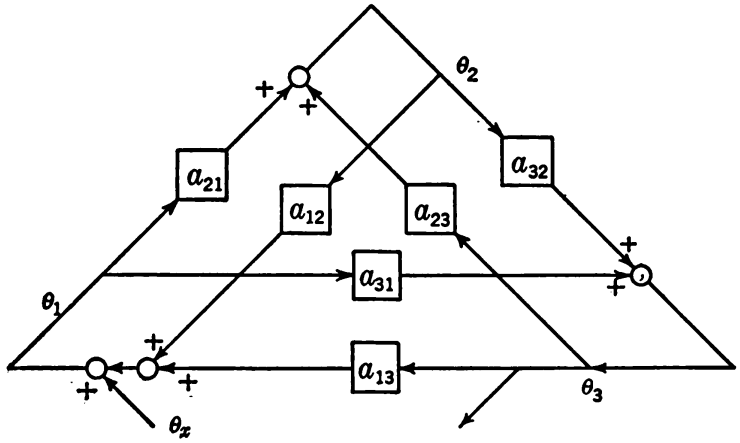


FIG. 8.32. Modification of Fig. 8.31 to use general transfer functions, in which the subscript identifies the input and output signals.

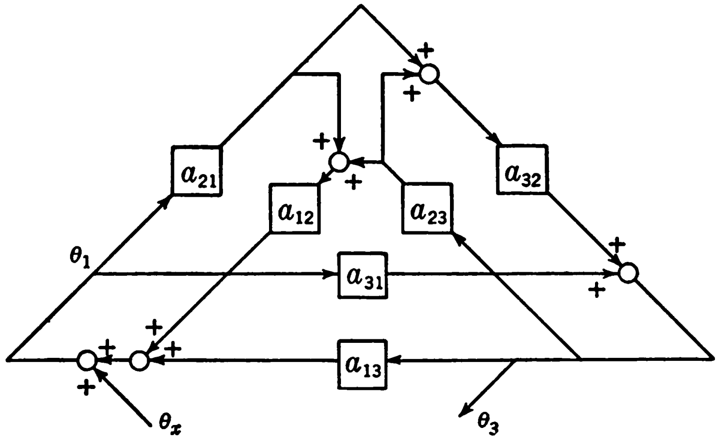


FIG. 8.33. Modification of Fig. 8.32 to eliminate θ_2 as a variable.

This technique could, of course, be extended to a system of any degree of complexity, but the problem of organizing the step-by-step solution of a complex system is so significant that other methods should be considered. The general problem is the solution of a set of simultaneous equations. The particular aspect of that problem being investigated here is the application of root-locus plots successively to solve the problem. Note that all that one plot can do is to find the values of s which make the sum of two terms equal to 0. Each term may be of any degree in s as long as its numerator and denominator are factored. A large number of plots, of course, are apt to be involved; the saving feature is that often only an approximate answer is needed and each plot can be sketched quickly. Finally, when accuracy is needed, the rough plots should indicate the parts of each plot which are critical in establishing the final plot.

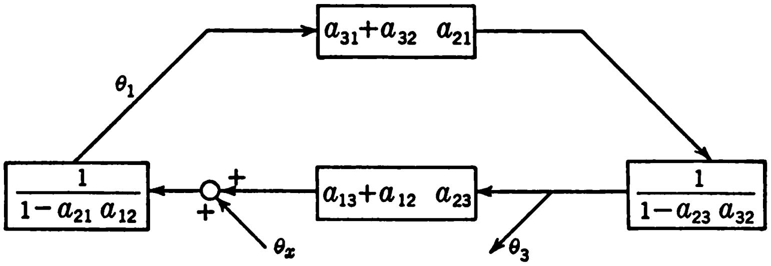


FIG. 8.34. Simplification of Fig. 8.33 to a single loop by combining inner loops and parallel paths by inspection.

In any case, however, the number of plots should be kept to a minimum and the significance of each one understood. Suppose, for example, that a complete system such as the pitch autopilot were analyzed and then the effect of some one parameter, such as the weight of the plane, was to be determined. This situation recalls a similar problem in circuit analysis, where the effect of varying a single resistor of a complicated network is involved. In this case Thévenin's theorem proves very effective by establishing the characteristics of the complete network as seen from the resistor.

Determinants provide an orderly procedure for the solution of simultaneous equations. A particular advantage is that the significance of any function is specified just by its position in a determinant. This simplification should help to catalogue the various problems.

The use of circuit-analysis concepts in conjunction with root-locus plots is presented in Chap. 9. The examples involve circuits which are frequently used in improvement of control systems. Determinants are considered in Chap. 10.

CHAPTER 9

CIRCUIT ANALYSIS

The representation of a system by a block diagram is very convenient for setting up root-locus plots when the variable is to be loop gain. Frequently, however, the loop gain is specified, and the optimum value of a circuit parameter, such as the capacitance in a lead network, must be selected. The circuit itself can be set up in block-diagram form, but the application of circuit-analysis concepts is suggested. Thévenin's theorem in particular proves to be a rapid means of setting up a root-locus plot with any circuit parameter as the variable. The theorem, which is easily proved by use of superposition, can be extended to apply to mechanical problems, such as the effect of compliance of a load shaft. A general procedure is established in which the zeros and poles for a root-locus plot are determined from the special cases when a parameter is zero and infinite. This technique is usually quite rapid and applicable to any system in which the parameter can be made to appear in only one equation.

9.1. Block Diagram of Circuits

> *A circuit can be set up in the form of a block diagram by treating impedances as transfer functions. The transfer function of a combination lag-lead circuit is then obtained by means of a root-locus plot using the long time constant as loop gain.*

Several circuits are used so frequently in control systems that they are practically standard. These circuits form a convenient set of examples in blending together concepts such as "impedance" in circuit analysis and "transfer function" in block diagrams. Start with the example of the integral network shown in Fig. 9.1a.

The transfer function from v_1 to v_2 is needed. The circuit-analysis solution involves dividing the input voltage v_1 by the total impedance of the loop to find the current i, and then multiplying by the impedance of the output branch to find the output voltage.

$$\begin{aligned}\frac{V_2}{V_1}(s) &= \left[\frac{I}{V_1}(s)\right]\left[\frac{V_2}{I}(s)\right]\\ &= \frac{1}{R_1 + R_2 + (1/C_2s)}\left(R_2 + \frac{1}{C_2s}\right) = \frac{1 + R_2C_2s}{1 + (R_1 + R_2)C_2s} \qquad (9.1)\end{aligned}$$

The impedance of the loop is the reciprocal of the transfer function from the applied voltage to the resultant current. The impedance of the output branch is the transfer function from current i to voltage v_2. The transfer function from v_1 to v_2 is then the product of the two transfer functions, as given in (9.1).

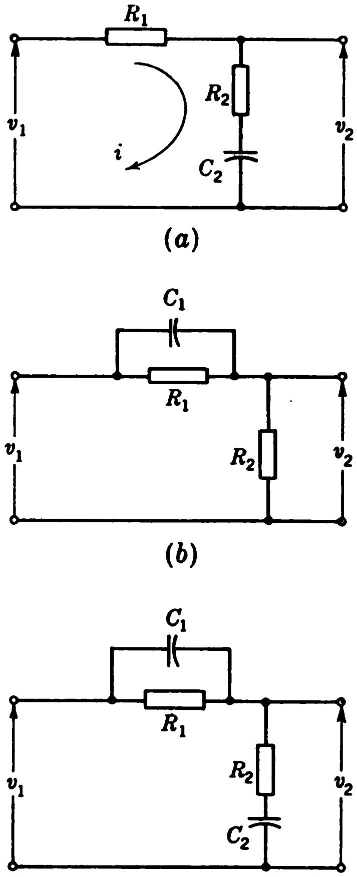
The lead circuit shown in Fig. 9.1*b* has the impedance of R_1 and C_1 in parallel treated as a single impedance, using a formula which is based on expressing the admittance of the parallel branches as the sum of the admittances of the individual branches.

$$Z_1 = \frac{1}{(1/R_1) + C_1s} = \frac{R_1}{1 + R_1C_1s} \qquad (9.2)$$

The transfer function from v_1 to v_2 can then be obtained in the same manner as in (9.1).

$$\begin{aligned} \frac{V_2}{V_1}(s) &= \frac{R_2}{R_2 + R_1/(1 + R_1C_1s)} \\ &= \frac{R_2(1 + R_1C_1s)}{R_1 + R_2 + R_1R_2C_1s} \qquad (9.3) \end{aligned}$$

The combination lag-lead network shown in Fig. 9.1*c* has a transfer function from v_1 to v_2 which could be obtained in the same manner. The algebra becomes somewhat involved, however, and the effect of individual parameters tends to become lost. Some interesting concepts are introduced in this example by solving the circuit in block-diagram form, as shown in Fig. 9.2.



(*a*)

(*b*)

(*c*)

FIG. 9.1. Circuits commonly used in control systems: (*a*) an integral network; (*b*) a lead network; (*c*) a combination integral and lead network.

The difference between the input voltage v_1 and the output voltage v_2 is the "error voltage" v_3. This voltage acting across R_1 and C_1 is multiplied by the sum of the admittances to obtain the total current i. This current times the impedance of the output branch is the output voltage

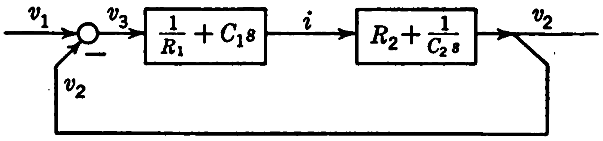


FIG. 9.2. Block diagram of circuit in Fig. 9.1*c* using voltages and currents as signals and impedances as transfer functions.

v_2. The roots of the characteristic equation of the complete loop are the values of s which make the transfer function around the loop equal to -1:

$$\frac{1 + R_1C_1s}{R_1}\,\frac{1 + R_2C_2s}{C_2s} = -1 \tag{9.4}$$

The root-locus plot versus $1/R_1C_2$ as "loop gain" is shown in Fig. 9.3. For low values of gain the root r_1 is essentially equal to $-1/R_1C_2$. Once a value of R_1C_2 is selected, the position of r_2 is automatically fixed. The transfer function from v_1 to v_2 is then

$$\frac{V_2}{V_1}(s) = \frac{(1 + R_1C_1s)(1 + R_2C_2s)}{(1 - s/r_1)(1 - s/r_2)} \tag{9.5}$$

The lead terms are present in this function because these terms were present in the transfer function of the forward path (Fig. 9.2).

Reviewing the transfer functions of these circuits reveals a characteristic which is common to them all and should provide a means of simplifying the analysis. Thus, in (9.1) the pole $-1/(R_1 + R_2)C_2$ corresponds to a transient exponential decay of C_2 through R_1 and R_2 in series, with the input short-circuited. Similarly, in (9.3) the pole $-(R_1 + R_2)/R_1R_2C_1$ corresponds to the decay of C_1 through R_1 and R_2 in parallel. Such values might be obtained directly from inspection of the circuit, as will be investigated in the next article.

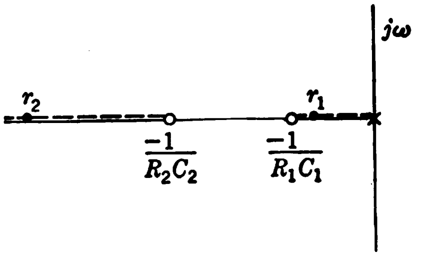


FIG. 9.3. Root-locus plot of system in Fig. 9.2 versus $1/R_1C_2$ as loop gain.

9.2. Zeros and Poles by Inspection

> *A pole of a transfer function corresponds to a transient of the form* e^{p_1t} *which can exist in the output with no input. A zero corresponds to a voltage of the form* e^{q_1t} *which can be impressed at the input with no voltage of that form appearing at the output. The poles and zeros can be determined by inspection for a system having a single energy storage.*

The three circuits studied in the previous article have transfer functions $(V_1/V_2)(s)$ as given in (9.1), (9.3), and (9.5), respectively. In the first two circuits only one capacitor is involved, and the pole in each case corresponds to the exponential decay of that capacitor with the input voltage source treated as a short circuit. In the third circuit, two capacitors are involved, and the poles are obtained by making the transfer function around the loop in Fig. 9.2 equal to -1. This condition makes

$v_2 + v_3 = 0$ and thus $v_1 = 0$; therefore, the poles again correspond to transients which can exist with the input short-circuited.

The zeros are also of the form $-1/RC$ for each circuit. In Fig. 9.1*a*, the zero corresponds to the output branch being short-circuited, with the transient being the decay of C_2 through R_2. In Fig. 9.1*b* the zero corresponds to the decay of C_1 through R_1, which would occur if the input were treated as an open circuit. Finally, in Fig. 9.1*c*, the zeros correspond to the two cases just considered. The consistency of these results suggests that obtaining the zeros and poles by direct inspection of a circuit would be a very useful procedure if it were thoroughly understood.

The significance of zeros and poles can be appreciated by study of the differential equations. Recall that a transfer function is obtained from a differential equation by substituting *s* for the operation *d/dt* (Chap. 4). The reverse substitution applied to (9.1) gives

$$\left(1 + R_2C_2\frac{d}{dt}\right)v_1 = \left[1 + (R_1 + R_2)C_2\frac{d}{dt}\right]v_2 \qquad (9.6)$$

The output v_2 is the sum of a steady-state and a transient response. The steady-state output is of the same form as the input voltage v_1, but the transient is the sum of the terms of the form e^{st} which make the right side of (9.6) equal to zero. The value $s = -1(R_1 + R_2)C_2$ satisfies this condition by making the total iR drop equal to the voltage across the capacitor at all times. The input is treated as a short circuit because it is a fixed connection in the circuit whose internal resistance has been assumed to be zero. The poles of a transfer function are the roots of the characteristic equation and correspond to transients of the form e^{st} which can exist in the output with no input function of that form. Thus, if v_2 were equal to e^{-t/R_2C_2} in (9.6), the left side of the equation would be zero. Inspection of Fig. 9.1*a* shows that a current of the form e^{-t/R_2C_2} flowing around the loop causes the output voltage v_2 to be zero because the sum of the iR_2 drop and the voltage across the capacitor C_2 is zero. *No short circuiting of the output branch* is involved; the path for the current is completed through the original circuit. The zero $-1/R_1C_1$ noted in (9.3) which arises for the circuit in Fig. 9.1*b* corresponds to the decay of C_1 through R_1. Again if the input voltage happened to be of the form e^{-t/R_1C_1} and of the same magnitude as the drop across R_1, the current flow through the main loop would be zero. The output voltage can then be zero because the input is balanced by an equal and opposite transient elsewhere in the circuit. The same type of reasoning used in these examples serves to explain the zeros and poles of the transfer function of the combination circuit.

The procedure in establishing the transfer function by direct inspection of a circuit can now be shown for a new circuit which involves a resistance load across the output of an integral network, as shown in Fig. 9.4.

The transfer function from v_1 to v_2, by direct inspection, is

$$\frac{V_2}{V_1}(s) = \frac{R_3}{R_1 + R_3} \frac{1 + R_2C_2s}{1 + \{R_2 + [R_1R_3/(R_1 + R_3)]\}C_2s} \tag{9.7}$$

The d-c value of the function is the simple resistance ratio because the capacitor C_2 then appears to be an open circuit. The zero is the same value $-1/R_2C_2$ as before, because with v_2 zero for this form of signal the presence of R_3 has no effect. The pole, however, is modified, because the decay of voltage from C_2 is now through R_2 and then R_3 in parallel with R_1.

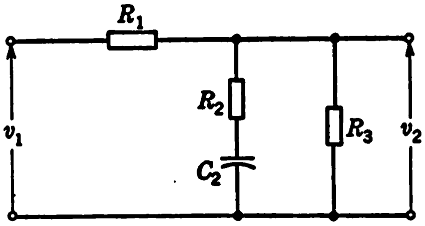


FIG. 9.4. An integral circuit with resistive load.

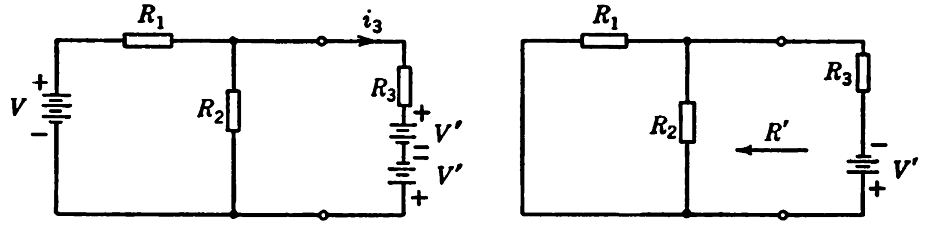
This method is very effective in systems in which only one energy storage is involved. When two or more forms of energy storage are involved, as in Fig. 9.1*c*, a root-locus plot will in general be needed to find the roots of the characteristic equation of the circuit. Thus far, a block diagram is first set up to determine the function to be used in making the plot. A method of setting up the plot directly from the circuit would be useful, particularly if any circuit parameter could be used as the variable. This procedure suggests the use of Thévenin's theorem, which replaces an entire network by a single loop in order that the effect of one parameter may be studied.

9.3. Use of Thévenin's Theorem

> *Thévenin's theorem states that, for the purpose of finding the current in a single branch, the remainder of the circuit can be replaced by the open-circuit voltage and the impedance seen from the branch. This theorem, which is easily proved, suggests a means of setting up a root-locus plot versus the value of the parameter in the branch.*

The proof of Thévenin's theorem is more valuable than the theorem itself. Consider the circuit shown in Fig. 9.5*a*: two batteries of equal voltage V' but of opposite sense are introduced into the branch in which the current is to be found. At first consider only the input battery voltage V and the top battery voltage V' to be acting in the circuit. If the voltage V' is made equal to the open-circuit voltage which would exist across the output terminals, then no current will flow through the load branch. Now consider the effect of the lower battery V' acting alone; the other two batteries have already been considered and are replaced by short circuits (Fig. 9.5*b*). The current produced in R_3 by the remaining battery is limited only by the resistance R_3 itself and the resistance of the rest of the circuit as measured between the output terminals with R_3 open.

The reasoning used in proving the theorem for the simple circuit above can be readily extended to apply to the circuit problem shown in Fig. 9.6. The problem is to find the effect of changing the load resistance R_3 upon the transfer function from v_1 to v_2 of the combination lag-lead network. The input voltage v_1 may be any arbitrary function of time, and the voltage v' inserted in the load branch may also be considered a function of time, with its instantaneous value being such that the current is zero in the load branch. The determination of this voltage v' due to the input voltage v_1 is then exactly the same problem that was solved for Fig. 9.1*c* with the result given in Eq. (9.5). Thus $(V'/V_1)(s)$ for the system in Fig. 9.6 is equal to $(V_2/V_1)\,(s)$ for the system in Fig. 9.1*c*, as given in (9.5), because the R_3 branch is open in both cases.



(*a*) (*b*)

FIG. 9.5. A justification of Thévenin's theorem: (*a*) a d-c circuit with equal and opposite batteries inserted in the load branch; (*b*) circuit after action of all but one battery is considered.

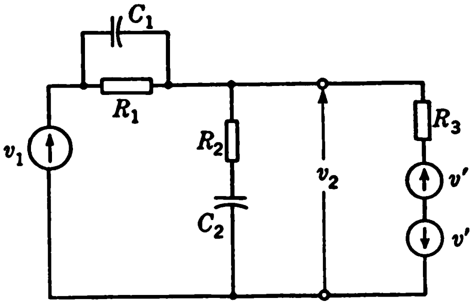


FIG. 9.6. The circuit of Fig. 9.1*c* with equal and opposite variable voltages inserted in the load branch.

The only current flowing in the resistor R_3 is then that due to the opposite voltage v' acting down. The impedance Z' of the original circuit as seen from the output terminals could be found in the usual manner of circuit analysis, but it is useful to determine this impedance by inspection, knowing that it must be of the form

$$Z' = Z_{dc}\frac{(1 - s/q_1)(1 - s/q_2)}{(1 - s/p_1)(1 - s/p_2)} \tag{9.8}$$

The d-c value of this impedance is simply R_1, because the capacitors are then open circuits. The zeros such as q_1 correspond to currents of the form e^{q_1t} which can flow into the impedance with the voltage across the impedance being zero. The only zero is then $q_1 = -1/R_2C_2$ corresponding to current e^{-t/R_2C_2} flowing through R_2 and C_2. The poles such as p_1 correspond to voltages of the form e^{p_1t} which could be impressed across the impedance without any current flowing in. These poles are the roots determined in Fig. 9.3 and are the same as the poles of the transfer function given in Eq. (9.5). The poles of the impedance function are always the same as the poles of the open-circuit transfer function because exactly the same circuit is involved; the input voltage is zero, and transients in the output voltage are being considered. The impedance looking into the circuit from the output terminals, given in (9.8), can thus be specified to be

$$Z' = R_1 \frac{1 + R_2C_2s}{(1 - s/r_1)(1 - s/r_2)} \tag{9.9}$$

The justification of this determination of the impedance by inspection is the same as that given in Art. 9.2 for the transfer function and is based on the meaning of the differential equation. An interesting check on the value of the impedance at very high frequencies can be obtained by simplification of (9.9) by approximating $1 + R_2C_2s$ as R_2C_2s, etc., to give

$$[Z']_{s\gg 1} = R_1 \frac{R_2C_2s}{(s/r_1)(s/r_2)} = \frac{R_1R_2C_2s}{R_1C_1R_2C_2s^2} = \frac{1}{C_1s} \tag{9.10}$$

The product $(1/r_1)(1/r_2)$ must be equal to $R_1C_1R_2C_2$, because the transfer function in (9.5) must be 1 for very high frequency signals. The final result $1/C_1s$ checks the circuit in that the impedance of the capacitor C_1 is lower than that of any other branch for high-frequency signals.

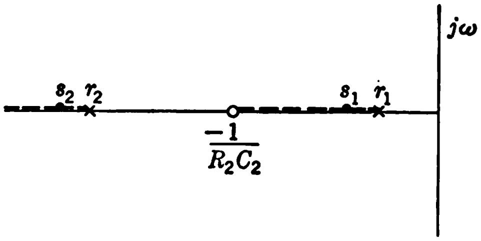


FIG. 9.7. Root-locus plot of system in Fig. 9.6 with R_3 variable.

Using Thévenin's theorem, the transfer function from v_1 to v_2 in Fig. 9.6 can now be expressed as

$$\frac{V_2}{V_1}(s) = \frac{(V'/V_1)(s)}{Z'/R_3 + 1} \tag{9.11}$$

The poles of this function are the roots of the equation $Z'/R_3 = -1$; these are determined in Fig. 9.7 and denoted as s_1 and s_2. The function $(V'/V_1)(s)$ for the system in Fig. 9.6 is equal to $(V_2/V_1)(s)$ for the system in Fig. 9.1*c*, given in (9.5). Substitution of this function and of Z' from (9.9) into (9.11) would show that the zeros of (9.11) are $-1/R_1C_1$ and $-1/R_2C_2$. These zeros could also be stated by inspection of Fig. 9.6 because the presence of R_3 does not change the arguments used to find the zeros of the system in Fig. 9.1*c*. The value of (9.11) for d-c signals, *i.e.*, for $s = 0$, is $1/(1 + R_1/R_3)$, making the full expression for (9.11) become

$$\frac{V_2}{V_1}(s) = \frac{(1 + R_1C_1s)(1 + R_2C_2s)}{(1 - s/s_1)(1 - s/s_2)(1 + R_1/R_3)} \tag{9.12}$$

9.4. Impedance Characteristics of Passive Networks

The impedance function of an RC network must have alternate zeros and poles along the real axis, with a pole being nearest to the origin, if the addition of another branch is not to lead to oscillatory roots.

Networks are a very convenient way to introduce new transfer functions into control systems, and it is well to know their characteristics. In particular, a passive network has no amplifier or other energy source, and therefore limitations can be specified independent of the arrangement of elements used. Networks involving only resistors and capacitors are particularly common, because inductances are heavier, subject to saturation, and not so readily available. Consider a general network with a resistance in one of the branches being variable, as shown in Fig. 9.8.

The purpose of this study is to determine as much as possible about the characteristics of the impedance Z as seen from the terminals outside the dashed-line box. Thévenin's theorem is not needed; the analysis can start from the voltage equation around the loop, treating Z as a single impedance.

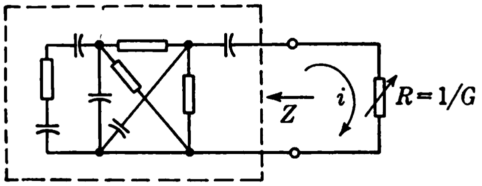


FIG. 9.8. A general network of resistances and capacitors.

$$iZ + iR = 0 \qquad \text{or} \qquad Z + R = 0 \tag{9.13}$$

The roots of the characteristic equation of this network are then the values of s which make the sum of the impedances zero or make the ratio of the impedances equal to -1.

$$Z\left(\frac{1}{R}\right) = ZG = -1 \tag{9.14}$$

Now Z is a ratio of polynomials both the numerator and the denominator of which can be expressed in factored form, even though the values of these zeros and poles are not as yet known. Writing Z in the general factored form, we have

$$ZG = Z_{dc}\,\frac{(1 - s/q_1)(1 - s/q_2)\cdots}{(1 - s/p_1)(1 - s/p_2)\cdots}\,G = -1 \tag{9.15}$$

Consider now that the admittance G of the external branch is changed gradually from zero to infinity, corresponding to a change from an open-circuit load to a short-circuit load. For a very low value of admittance, the natural transients of the circuit are essentially the same as when the external branch is open. But note in (9.15) that, for very small values of G, the impedance Z must be very large in order that the product ZG may be 1. For Z to be very large, the value of s must be nearly the same as one of the poles, such as p_1, in order that a denominator term may be very small. This checks the previous result that poles of the impedance function are the roots of the characteristic equation of a circuit.

At the other extreme, where G is very large, Z must be very small. Thus, s must be nearly equal to one of the zeros in order that a numerator term may be very small. The zeros are the roots corresponding to the case when the external branch is a short circuit. Another possibility for a root exists in both cases, depending upon whether the numerator or the denominator is of higher order in s. If the denominator is of the fourth degree, for example, with the numerator of third degree, then the impedance can be made zero by an infinite value of s.

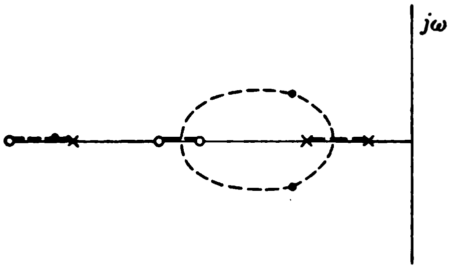


FIG. 9.9. A root-locus plot versus R for the circuit in Fig. 9.8, assuming random sequence of zeros and poles of the input impedance function.

The roots correspond to transients of the form e^{rt} which can exist in the transient response of the network. Such a transient could be started by charging up one capacitor with all others discharged. The one capacitor would then discharge through resistors into other capacitors, dissipating its stored energy either as heat in the resistors or by storing the energy in other capacitors. But the discharge must proceed with the current in any branch always being from a point of high potential to a point of lower potential. The voltage on the capacitor which had been charged will thus always decrease, never going negative or oscillating. Thus, all the roots of the system must be negative real numbers, because complex ones would correspond to oscillations. This restriction on the roots is true regardless of whether the external branch is open circuit or short circuit, and thus all zeros and poles of Z must lie along the negative real axis.

The roots must also always lie along this axis for any intermediate value of G. The locus of roots versus R for an arbitrary sequence of zeros and poles, however, would be as shown in Fig. 9.9. Two zeros or two poles being adjacent to each other permits the locus to break away from the real axis; therefore, the zeros and poles must alternate. The only remaining question is whether the point nearest the origin should be a zero or a pole. To establish this point, consider that the external branch is a capacitor; then G in (9.14) is replaced by Cs, which introduces a zero at the origin. Therefore, the first point of the impedance must be a pole if the alternation of poles and zeros is to be maintained.

The same type of reasoning can be applied to a general circuit containing capacitors and pure inductances. The limitations in this case are described by Foster's theorem; root-locus plots provide an interesting demonstration. Foster's theorem states that the zeros and poles are conjugate pairs which alternate along the $j\omega$ axis. If the pair nearest the origin are poles, then a zero exists at the origin; if the first pair are zeros, then a pole exists at the origin. Consider the addition of either a capacitor or an inductance whose value may be varied from zero to infinity. The locus must now lie along the $j\omega$ axis because the circuit elements can neither dissipate nor supply energy; oscillations once started continue at constant amplitude. The admittance of the capacitor puts a zero at the origin, and that of the inductance puts a pole at the origin. In any case, however, the product has either no zeros or poles at the origin or a pair of one of them; thus, the locus from the origin moves along the $j\omega$ axis also.

For a general circuit containing resistance as well as inductance and capacitance, the restrictions are not easy to specify, because the zeros and poles can then be anywhere in the left half plane. A single example, however, shows the effect of adding another branch.

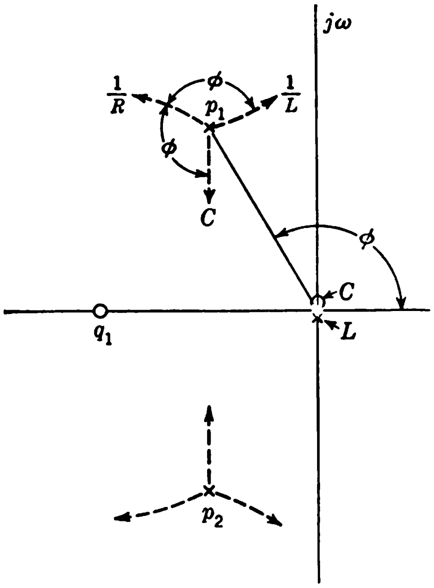


FIG. 9.10. Root-locus plot for the addition of resistance, capacitance, or inductance across a circuit having an input impedance $R(1 - s/q_1)/(1 - s/p_1)(1 - s/p_2)$.

The impedance Z as viewed from a given pair of terminals is assumed to be

$$Z = \frac{R(1 - s/q_1)}{(1 - s/p_1)(1 - s/p_2)} \tag{9.16}$$

The admittance of the added branch may be either $1/R$, Cs, or $1/Ls$ depending upon the element added. The initial direction of the locus in each case is shown in Fig. 9.10.

For resistance added, no pole or zero is present in the admittance G; the locus starts from the pole p_1 in a circle about q_1 as a center. For a capacitor added instead, a zero is added at the origin, and the locus starts from p_1 in a direction which is counterclockwise from the previous locus by the angle $\phi°$. Similarly the connection of the inductance adds a pole at the origin, and the locus starts in a direction which is $\phi°$ clockwise from the locus versus $1/R$.

9.5. Analysis of a Bridged-T Network

The zeros of a bridged-T network can be determined by a root-locus type of plot. For a very low damping ratio of the zeros, the resistances should be equal and the ratio of capacitors should be small.

A modulated carrier signal arises in control systems when a synchro, excited at a carrier frequency of, say, 400 cps, is modulated by mechanical oscillation of the rotor at a low frequency, such as 5 cps. A bridged-T network is used to achieve the same effect with this type of signal as that produced by a simple lead network with a d-c potentiometer signal. It will be shown in the next article that the bridged-T network characteristic versus frequency must have a very low output at the carrier frequency and a rapid increase in output on either side of this frequency point. Such a characteristic requires that a zero of the transfer function be close to the $j\omega$ axis at the carrier frequency.

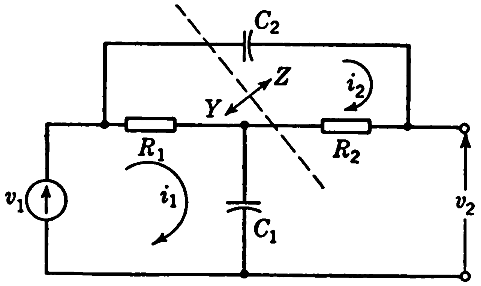


FIG. 9.11. A bridged-T network divided for convenience in finding roots.

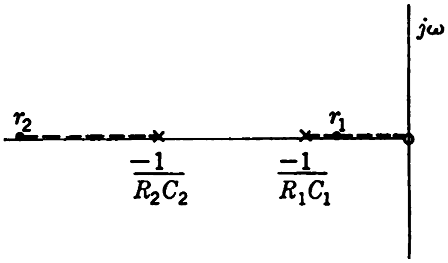


FIG. 9.12. Root-locus plot for circuit in Fig. 9.11 with $1/R_1C_2$ as the variable.

First consider the problem of finding the poles of this transfer function. Using the concepts of the previous article, the network may be divided into two sections, each of which has an impedance as viewed from the two terminals of the break. The poles are then the values of s which make the sum of these two impedances equal to zero or make their ratio equal to -1. A very convenient way to divide the circuit is shown by the dashed line in Fig. 9.11.

The product of the impedance of the upper section times the admittance of the other section becomes simply

$$ZY = \left(R_2 + \frac{1}{C_2 s}\right)\left(\frac{1}{R_1} + C_1 s\right) \tag{9.17}$$

The roots of the network equation are then the values of s which make this product or $1/ZY$ equal to -1, as shown in Fig. 9.12.

The next problem is to find the zeros of the transfer function; these are values of s corresponding to transients of the form e^{st} which can exist in the input with no output. The equations will be set up for the circuit according to the loop currents as shown in Fig. 9.11.

The summation of voltages around the C_2 loop is

$$i_1R_1 = i_2\left(R_1 + R_2 + \frac{1}{C_2s}\right) \tag{9.18}$$

The condition for zero output voltage can be expressed as

$$i_1\frac{1}{C_1s} + i_2R_2 = 0 \tag{9.19}$$

No equation is needed involving v_1 because v_1 can be whatever is necessary to make the currents flow with no output voltage. Solve for i_1 from (9.19) and substitute it into (9.18):

$$R_2C_1s + \frac{1 + (R_1 + R_2)C_2s}{R_1C_2s} = 0 \tag{9.20}$$

The roots of (9.20) are then the values of s which make the ratio of these two functions equal to -1, as shown in Fig. 9.13.

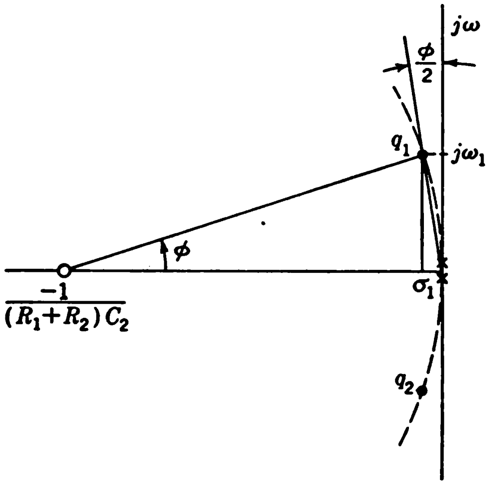


Fig. 9.13. Root-locus plot to determine the zeros of the transfer function of the circuit in Fig. 9.11.

This locus of roots is a circle about the point $-1/(R_1 + R_2)C_2$ as center, making $|1 + (R_1 + R_2)C_2s| = 1$, and thus

$$|q_1| \cong \frac{1}{\sqrt{R_1C_1R_2C_2}} \tag{9.21}$$

For small values of $\phi/2$, the damping ratio is

$$\zeta \cong \frac{\phi}{2} \cong \frac{1}{2}\,\omega_1(R_1 + R_2)C_2 \cong \frac{1}{2}\,\frac{R_1 + R_2}{\sqrt{R_1R_2}}\sqrt{\frac{C_2}{C_1}} \tag{9.22}$$

The damping ratio must be of the order of 0.05 in order for the network to be effective on the modulated carrier signal, as discussed in the next article. Therefore, the ratio of C_2/C_1 must be very small; the values R_1 and R_2 should be equal if the sum over the geometric mean of resistances is to be a minimum. The roots of (9.20) are the zeros of $(V_2/V_1)(s)$, making the complete transfer function for the circuit

$$\frac{V_2}{V_1}(s) = \frac{(1 - s/q_1)(1 - s/q_2)}{(1 - s/r_1)(1 - s/r_2)} \tag{9.23}$$

9.6. Modulated Carrier Signals

> *A modulated carrier signal can be resolved into two sine waves, having frequencies above and below the carrier frequency by the amount of the modulated frequency. These signals are transmitted through the bridge network and then recombined into a modulated carrier. A vector construction shows that a zero close to the $j\omega$ axis at the carrier frequency produces phase lead for the modulation envelope.*

A synchro transmitter, having a constant angular error, will give an a-c signal of the frequency of the excitation of the synchros, $\omega_c/2\pi = 400$ cps. If the angular error itself is a sine wave of a much lower frequency, $\omega_m/2\pi$ cps, the instantaneous value of the error-voltage signal in terms of vector lengths in Fig. 9.14 will be

$$OP = OM \cos \omega_c t \cos \omega_m t \qquad (9.24)$$

The terms "carrier frequency" ω_c and "modulation frequency" ω_m are used because the same type of signal is common to radio circuits.

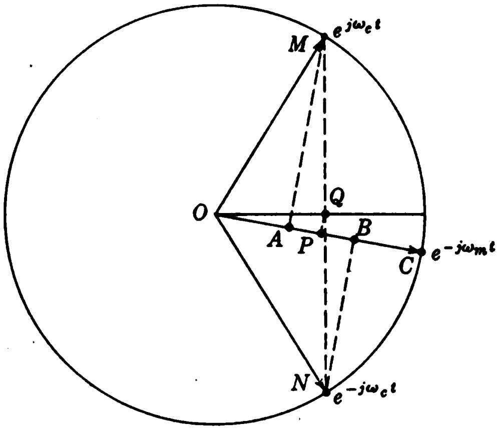


FIG. 9.14. A vector diagram to demonstrate the equivalence of a modulated carrier signal and two signals with offset frequencies.

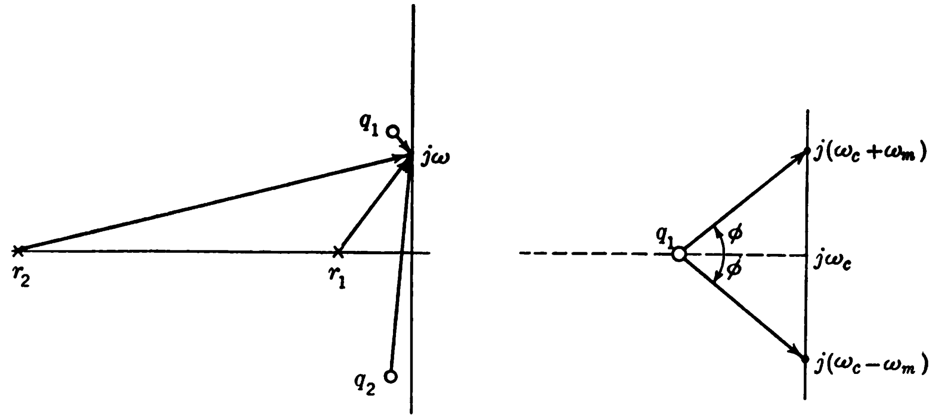
The instantaneous voltage can be represented vectorially as shown in Fig. 9.14. The vector $e^{j\omega_c t}$ rotates at a circular frequency ω_c; its projection on the real axis OQ is $\cos \omega_c t$. The vector $e^{-j\omega_m t}$ rotates at angular speed ω_m; the projection of $\cos \omega_c t$ on it, OP, is $\cos \omega_c t \cos \omega_m t$. The same length OP can be obtained as the average of projections OA, from $e^{j\omega_c t}$, and $e^{-j\omega_c t}$, directly on $e^{-j\omega_m t}$. The proof is that the line lengths MQ and QN are equal by symmetry; the projections AP and PB of these line lengths on the line OC are also equal. The instantaneous value of the voltage, OP, is then the average of the two values OA and OB.

The geometry of obtaining OA and OB is exactly the same, however, if the reference line OC is held fixed at the real axis and the top vector made to rotate at the rate $\omega_c + \omega_m$ and the bottom vector made to rotate at the rate $\omega_c - \omega_m$.

$$\cos \omega_c t \cos \omega_m t = \tfrac{1}{2}[\cos (\omega_c + \omega_m)t + \cos (\omega_c - \omega_m)t] \qquad (9.25)$$

This vectorial conversion can also be carried out by means of trigonometric identities. The significant thing is that the modulated carrier signal is exactly equal to the sum of two sinusoidal signals having frequencies which are the sum and difference of the carrier and modulation frequencies.

When the modulated carrier signal is impressed at the input of the bridged-T network, the output can be found by considering the signal to be the sum of the two beat-frequency signals. The value of the transfer function can be easily estimated from the complex-plane plot shown in Fig. 9.15. A general value of the complex number s is shown. For s equal to zero the phase shift of the bridge is zero, and the attenuation ratio is unity. As s moves up the $j\omega$ axis, all the vectors in the transfer function follow the s point, pivoting at the poles and zeros. In the region of $j\omega_c$ the effect of all vectors is nearly constant except for that of the one pivoted at the zero q_1. A close-up of this region, shown in Fig. 9.15*b*, indicates that the higher-frequency signal is given a phase shift by an angle ϕ more than that of a signal at carrier frequency and the lower-frequency signal is given phase-shift lag by the angle ϕ.



(*a*) (*b*)

FIG. 9.15. Transmission of signals through the circuit in Fig. 9.11: (*a*) vectors from all poles and zeros of the transfer function; (*b*) expanded view of the effect of the zero q_1 near the $j\omega$ axis.

The rotating-vector picture for the output signals of the bridged-T network is shown in Fig. 9.16. The phase lag caused by the poles at r_1 and r_2 and the zero at q_2 rotates each vector back toward the time axis, OM to OM' and ON to ON', as shown in Fig. 9.16*a*. The average of the projections OA and OB upon the time axis can be resolved into a single vector OP, obtained by projection of OM' to OQ and then to OP, to give a modulated carrier signal again. This effect is to shift the phase of the carrier only, not of the modulation envelope.

The opposite phase shifts ϕ due to the zero q_1, however, require that the top vector OM be rotated away from the time axis and the bottom vector ON be rotated toward the time axis. The same relative shift can be achieved simply by rotating the time axis OC through the angle ϕ to OC', as shown in Fig. 9.16*b*. Again the same construction as in Fig. 9.16*a* obtains OP as the average of OA and OB, to give a modulated carrier signal. The effect of the zero q_1, however, is to shift the phase angle of the modulation envelope ahead by the angle ϕ without affecting the carrier. This shift in the envelope is the lead angle ϕ desired and is equal to $\tan^{-1} \omega_m/\zeta\omega_c$, in which ζ is the damping ratio of q_1.

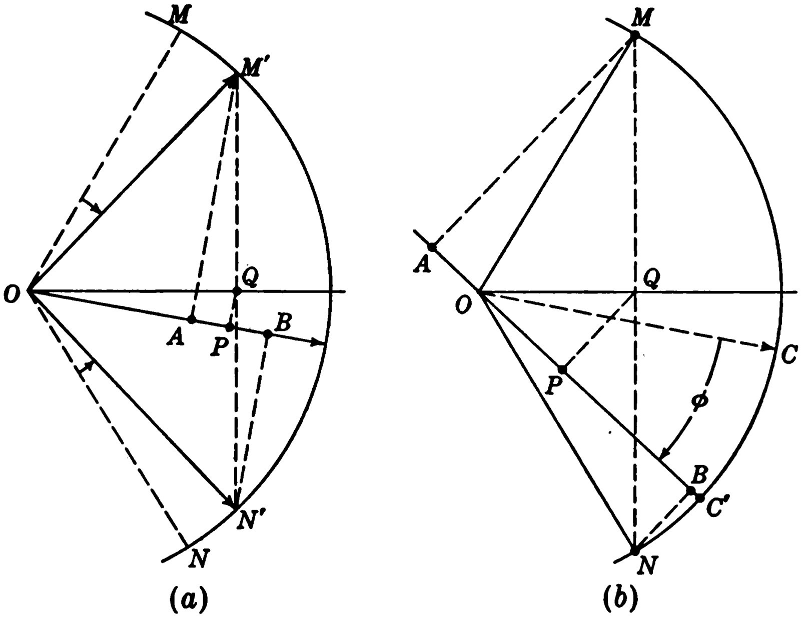


FIG. 9.16. Effect of bridged-T network on a modulated carrier signal: (*a*) equal phase shift of both signals; (*b*) opposite phase shifts due to zero near to $j\omega$ axis.

9.7. Choice of Capacitor in a Complete System

> *The locus of roots can be plotted versus capacitance if the rest of the system is treated as an impedance. The poles of the impedance function are the roots of the system with capacitance zero; the zeros are the roots with capacitance infinite. These extreme conditions correspond to open circuit and short circuit, for which the roots are already known from a plot versus loop gain. The subsequent locus versus C permits the direct choice of C for maximum damping of the complex roots.*

Circuit concepts, such as input impedance, can be applied to the problem of selecting the capacitor to be used in a lead network. The standard system (Fig. 5.1) is represented in Fig. 9.17 in a form which is partially circuit and partially block diagram.

The entire control system is shown enclosed in a dashed block, because it will be treated as a single impedance as viewed from the terminals of the capacitor. This impedance is the ratio of the impressed voltage to resultant current. The d-c value of impedance involves replacement of the capacitor by a d-c battery; current will flow around the loop through R_1. If current flows through R_2, the voltage input to the amplifier will cause the motor to drive the output until the output of the first amplifier opposes the applied battery voltage and reduces the current through R_2 to zero. The d-c value of the impedance, as viewed from the terminals of the capacitor, is thus simply R_1.

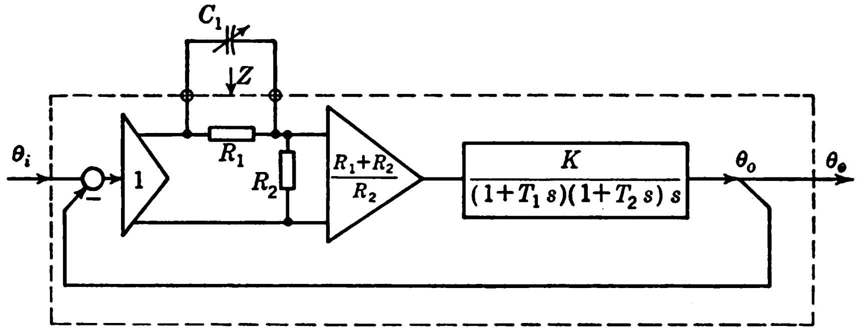


FIG. 9.17. The basic system in Fig. 5.1 modified to include a lead network and arranged to treat the entire system as an impedance viewed from the terminals of the capacitor.

The roots of the equation of the complete system are the values of s which make the impedance Z times the admittance of the capacitor, C_1s, equal to -1.

$$ZY = \frac{R_1(1 - s/q_1)(1 - s/q_2) \cdots}{(1 - s/p_1)(1 - s/p_2) \cdots} C_1 s = -1 \quad (9.26)$$

The zeros and poles of this impedance function can be determined by use of the same arguments as in Art. 9.4. For $C_1 = 0$ the roots of the system must be equal to the poles of Z, such as p_1; for $C_1 = \infty$, the roots of the system must be the zeros of Z, such as q_1. These conclusions are based on keeping the magnitude of ZY equal to 1 as the value of C_1 is allowed to approach 0 and ∞, respectively. But for $C_1 = 0$, the system is the original system, whose roots have been determined by a locus plot in Fig. 7.8 with loop gain K_2 and are denoted by the points p_1, p_2, and p_3 in Fig. 9.18.

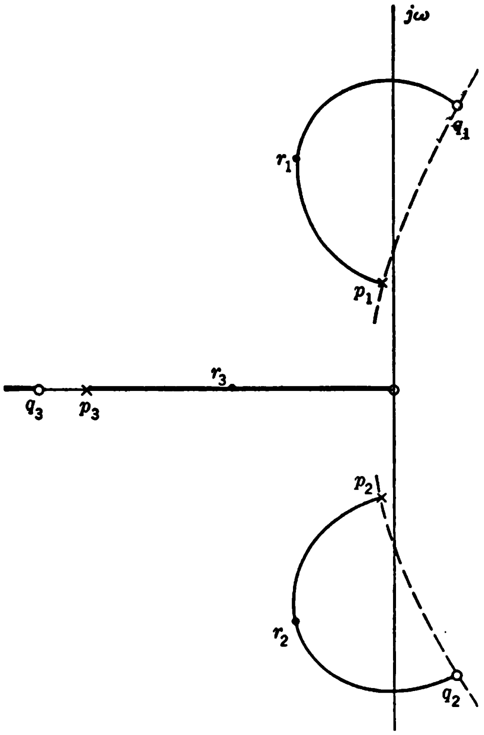


FIG. 9.18. Root-locus plot for system in Fig. 9.17. With C_1 variable, solid lines; versus K, dashed lines.

Digitized by Google Original from UNIVERSITY OF MICHIGAN

The attenuation of the resistance net is canceled by an equal amplifier gain, so that the net d-c gain of the loop, K, is not changed. At the other extreme, $C_1 = \infty$, the voltage across C_1 will always be zero; therefore, the attenuation of the resistance network is eliminated. The loop gain is thus increased by the factor $(R_1 + R_2)/R_2$, and the roots are farther out along the dashed locus, as shown by the points q_1, q_2, and q_3.

In the solution of (9.26) the poles are the start of the locus versus C_1 and the zeros are the terminal points. Note that the admittance C_1s adds a zero at the origin. The locus versus C_1 is shown in Fig. 9.18 by solid lines.

Suppose that the capacitor C_1 were replaced by a variable resistance; in this case the admittance would have no zero introduced at the origin. Increasing G would simply increase the loop gain, and the locus would move out along the same path as that versus gain K. But the addition of the zero at the origin which arises for $Y = C_1s$ causes an additional phase lead to be present at p_1 equal to ϕ_0. The locus thus starts from p_1 at the angle ϕ_0 counterclockwise from the gain locus in order that a compensating lag angle ϕ_0 may be introduced by the vector $s - p_1$.

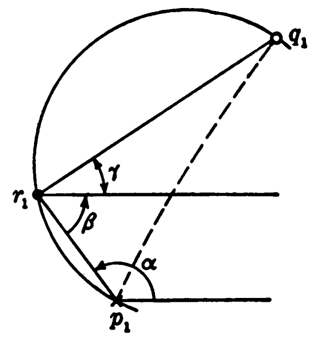


FIG. 9.19. Circular locus between a pole and a zero, assuming that net phase shift due to all other poles and zeros is a constant.

The locus from p_1 to q_1 is approximately an arc of a circle, as can be shown by assuming that only the vectors from the points p_1 and q_1 change appreciably (Fig. 9.19). The net angle of lead at r_1 due to the pole at p_1 and the zero at q_1 is $(\gamma + 180) - \alpha$. But $180 - \alpha$ is β; therefore the net lead angle is $\gamma + \beta$. The angle $\beta + \gamma$ is constant if r_1 lies along the arc of a circle between p_1 and q_1. The net phase angle of other vectors in Fig. 9.18 is not constant, however, causing the locus to be distorted from a circular shape.

The maximum damping constant of the complex roots r_1 and r_2 can be achieved at the point of tangency of the locus with a vertical line. The corresponding value of C_1 is then calculated from (9.26). As a first approximation, assume that the product of all terms involving the zeros and poles of Z is equal to $\frac{1}{3}$. The magnitude condition of (9.26) is then satisfied simply by making $R_1C_1 = 3/|r_1|$.

The roots of the system are the poles of the transfer function from θ_i to θ_o. The zeros of this function are the values of s corresponding to transients of the form e^{st} which can exist in the input θ_i with no output θ_o. Inspection of Fig. 9.12 reveals that the only zero is $-1/R_1C_1$, because an input voltage e^{-t/R_1C_1} could cancel a transient voltage across R_1, leaving the voltage across R_2 zero. The complete transfer function is, therefore,

$$\frac{\Theta_o}{\Theta_i}(s) = \frac{1 + R_1C_1s}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)(1 - s/r_4)} \tag{9.27}$$

The root r_4 is off scale on the plot.

9.8. Simultaneous Choice of Two Parameters

Addition of a shunt capacitance to a lead network can be studied by the technique of finding the roots for the open-circuit and short-circuit conditions. A grid of loci can be sketched in the complex plane to permit simultaneous selection of both values of capacitance.

The process described in Art. (9.7) can be used to include new changes in the system. Thus by adding a capacitor in series with R_2, a combination integral-lead network is obtained, as shown in Fig. 9.20.

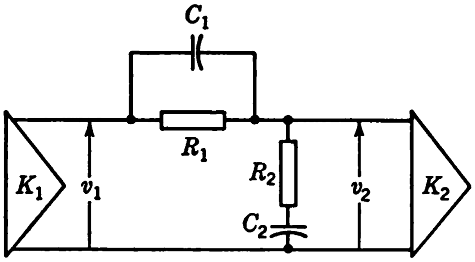


FIG. 9.20. A combination integral-lead network added to the basic system of Fig. 5.1.

For constant-velocity inputs, the steady-state error signal is not attenuated by a resistor divider as before. For $C_2 = \infty$, the circuit is the same as in Fig. 9.17, and the roots of the system are the values determined in Fig. 9.18. The root to the left of q_3 is neglected. At the other extreme, $C_2 = 0$, the attenuation of the resistance network is eliminated for all signals, and q_1, q_2, and q_3 in Fig. 9.18 become roots of the system. The equation for the roots of the system in terms of the admittance of the network, Y_2, and the impedance of the capacitor, $1/C_2s$, will be shown to be

$$Y_2\frac{1}{C_2s} = \frac{(1 - s/q_1)(1 - s/q_2)(1 - s/q_3)}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)}\frac{1 + R_1C_1s}{R_2}\frac{1}{C_2s} = -1 \tag{9.28}$$

The open-circuit roots are the zeros of the admittance function, and the short-circuit roots are the poles. The d-c value of the admittance can be shown to be $1/R_2$ by replacing the capacitor with a battery and noting that the steady-state current flow makes $v_2 = 0$. The new zero $-1/R_1C_1$ is due to the new transient which can exist with $C_2 = 0$ because a circulating current around the R_1C_1 loop cannot be influenced by any action of the servo.

The root-locus plot for (9.28) versus C_2 is shown in Fig. 9.21. The locus from r_1 now starts to the right and crosses the $j\omega$ axis quickly, making the allowable value of $1/R_2C_2$ very small. The change in the angle at which the locus emerges from r_1 compared to the direction of the locus in Fig. 9.18 is due to replacement of the zero at the origin by a pole, with the new zero appearing at $-1/R_1C_1$. A new root is between $-1/R_1C_1$ and the origin; analysis of the same result in Art. 5.10 shows that the low steady-state error allowed by C_2 is achieved slowly. Reducing C_2 is desirable to move this real root to the left, but some means of increasing the damping of the complex roots must first be found.

Reduction in gain achieves greater damping. The amplifier gain had been increased to compensate for the attenuation of the lead network; then the insertion of C_2 avoided this attenuation. The net result is that static gain is higher by the factor $(R_1 + R_2)/R_2$ than the originally specified value. A reduction of gain could be made as a new modification of the existing system, or the gain could be lowered in Fig. 9.18 and the sequence of plots resketched. This second sketching of the plots is faster than the first, because the zeros and poles involved in each plot, as well as the general shape of the loci, are already known. Moreover, the study of the simultaneous choice of both capacitors is aided by establishing a grid of loci across the region in which the roots must lie, as shown in Fig. 9.22.

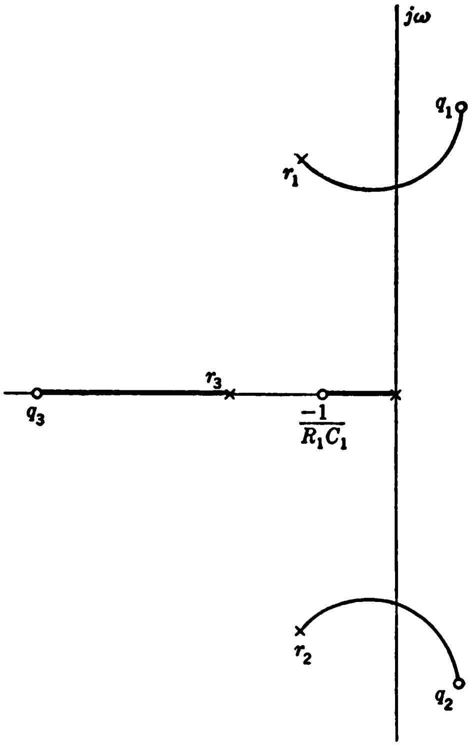


FIG. 9.21. Root locus of the system in Fig. 9.20 with C_1 fixed and C_2 variable.

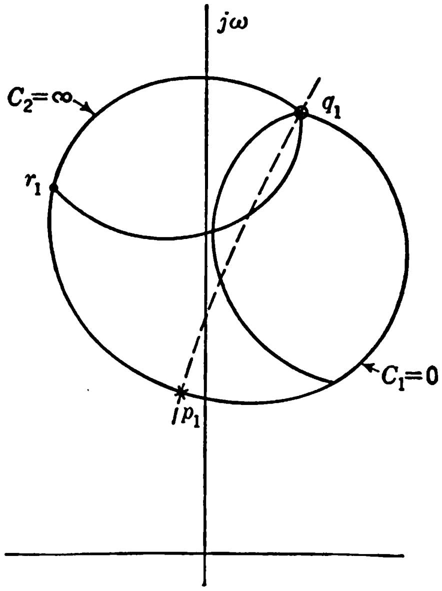


FIG. 9.22. Region of shift of the root r_1 due to variation of both C_1 and C_2.

The upper curve is the locus versus C_1 with C_2 infinite; the lower curve is the locus versus C_2 with C_1 zero. The locus from r_1 to q_1 is similar to the locus established in Fig. 9.21. Other loci can be sketched from one curve to the point q_1 in each case. This grid of lines aids in the simultaneous choice of C_1 and C_2.

9.9. Locus Versus a Mechanical Parameter

> *The technique of establishing the position of the roots for a parameter having the extreme values of zero and infinity can be used with mechanical parameters. When the compliance of a motor shaft is variable, the limiting cases are a solid shaft and a broken shaft.*

Consider the standard system (Fig. 4.1), but let compliance of the shaft between the motor and the load be varied, as shown in Fig. 9.23. This spring effect is present in an autopilot servo in the shaft connecting the actuator to the control surface. The effect is sufficiently serious that the actuator must often be mounted very close to the control surface, even though a simpler design could be made using a torsion tube between the actuator and the control surface.

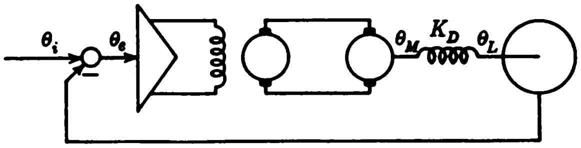


FIG. 9.23. The basic system of Fig. 4.1 with compliance of the shaft between motor and load added.

The application of Thévenin's theorem involved the concept of the impedance of the system as seen from the terminals of the variable element such as a capacitor. For this mechanical system the "mechanical impedance" of the system seen from the spring K_D is the ratio of the torque impressed on the system between the ends of the spring and the resultant angle $\theta_M - \theta_L$, in which θ_M is the motor angle and θ_L is the load angle. Denoting this impedance as $K_CG_C(s)$, the characteristic equation of the system is

$$\frac{K_CG_C(s)}{K_D} = -1 \tag{9.29}$$

Nothing has as yet been determined about $K_CG_C(s)$, but writing it in terms of its zeros and poles with the possibility of zeros or poles at the origin makes (9.29) become

$$\frac{K_Cs^\gamma(1 - s/q_1)(1 - s/q_2)\cdots}{(1 - s/p_1)(1 - s/p_2)\cdots}\,\frac{1}{K_D} = -1 \tag{9.30}$$

For K_D infinite, the system is the same as that in Fig. 5.1, for which the root-locus plot versus gain K is given in Fig. 7.8. Selecting $K = K_2$, the roots are r_1, r_2, and r_3, which are replotted from Fig. 7.8 in Fig. 9.24. For $K_D = 0$, the control loop is broken, and the roots of the system are the roots of the equations for the individual components. One root is $-1/T_1$ for the generator field; the other is $-1/T_2$ for motor acceleration, although the numerical value would be altered from that previously used because the inertia of the load is uncoupled from the motor. The poles and zeros of (9.30) are thus determined, but $K_C s^\gamma$ depends upon the steady angle $\theta_M - \theta_L$ produced by a steady-state torque between motor and load. This torque would act through the transfer function $1/J_L s^2$ to produce θ_L, which in turn acts through K_2/s to produce θ_M. The specific form of (9.30) is therefore

$$\frac{J_L s^2 s}{K_2} \frac{(1 + T_1 s)(1 + T_2 s)}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)} \frac{1}{K_D} = -1 \qquad (9.31)$$

The root-locus plot for (9.31) versus K_D as the variable is then as shown in Fig. 9.24. As the compliance of the load shaft increases from zero, the locus from the poles at r_1 and r_2 moves up, around to the left, and then back into the real axis. This locus corresponds to the servo roots returning to the open-loop poles as the loop is opened gradually by weakening the load shaft. A new pair of roots starts from the ends of the $j\omega$ axis; these correspond to the natural resonance of the shaft itself. As the shaft is weakened, these roots are moved into the right half plane by the action of the servo, as shown by the roots s_3 and s_4. For very high values of compliance, friction effects will no longer be negligible. For very low values of compliance, very short time constants not considered in the equation will influence the position of the locus with respect to the $j\omega$ axis. The limiting positions of the locus are thus convenient in establishing the plot, but the limits themselves should not be accepted unless the assumptions are still valid.

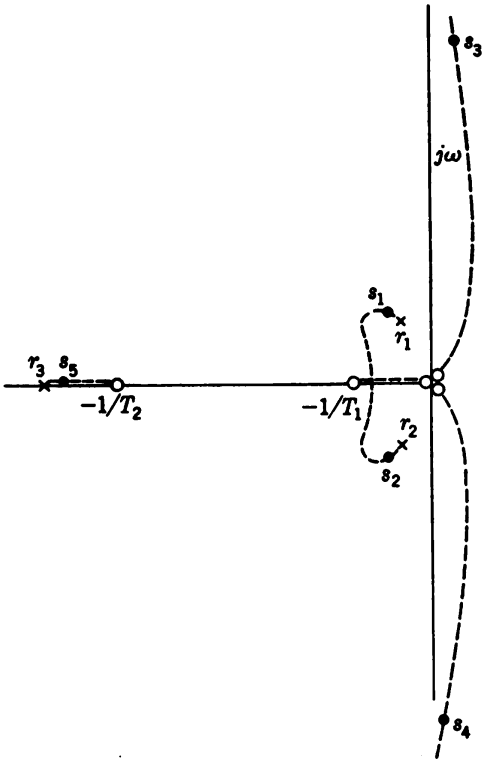


FIG. 9.24. Root-locus plot of system in Fig. 9.23 versus compliance of shaft, $1/K_D$.

9.10. Correlation with Other Methods

The technique of sketching a locus versus any variable can be justified using either block diagrams or determinants. The simplest test *of applicability of the technique is that a set of equations can be written describing the system in which the variable appears in only one equation.*

Any one of the systems solved in this chapter could have been set up and solved in block-diagram form. Thus the final problem of the previous article (Fig. 9.23) would appear in a block-diagram form as shown in Fig. 9.25.

This block diagram can be set up in simple form by assuming that the compliance of the load shaft is zero. Then include compliance by having the twist of the shaft subtracted from motor angle θ_M to produce the load angle θ_L. The load angle θ_L times $J_L s^2$ is the torque m_L which must exist in the spring. This torque multiplied by the compliance $1/K_D$ then determines the twist of the shaft. Compliance is thus obtained as a parameter which appears in only one loop. The rest of the system can be reduced to a single function as seen from this loop, and the final root-locus plot is made with compliance as the variable. These loci must be continuous curves which start from the points for compliance zero and end on the points for compliance infinite. These sets of points must be the poles and zeros of the transfer function of the system from $\theta_M - \theta_L$ to m_L, in order that these end conditions may be met. Being poles and zeros of the loop transfer function, these points must also be the base points for vectors which determine the position of the locus.' The significant point to note from the block-diagram reasoning is that the element in the system which is to be varied should be set up as an isolated parameter in a single loop.

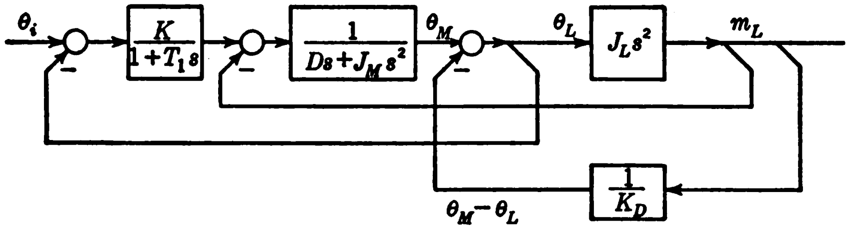


FIG. 9.25. A block diagram of the system in Fig. 9.23 with the compliance isolated.

An example where a parameter cannot be isolated occurs in the study of the effect of the angular momentum of a gyroscope on the roots of a system containing that gyroscope. The momentum will appear in two blocks of the block diagram, because deflection of the momentum vector in either of two directions is important.

The procedure of setting up a root-locus plot versus a variable in the system can also be justified by using some of the concepts usually associated with determinants. The arguments are quite simple and can be illustrated in terms of the equations directly. The system shown in block-diagram form in Fig. 9.25 will be used. Letting $\theta_i = \Theta_i e^{st}$, $\theta_e = \Theta_e e^{st}$, $m_L = M_L e^{st}$, $\theta_L = \Theta_L e^{st}$, and $\theta_M = \Theta_M e^{st}$, these equations become

$$\Theta_i = \Theta_e + \Theta_L \tag{9.32}$$

$$0 = \frac{K\Theta_e}{1 + T_1 s} - M_L - (Ds + J_M s^2)\Theta_M \tag{9.33}$$

$$0 = \Theta_L - \frac{1}{K_D} M_L + \Theta_M \tag{9.34}$$

$$0 = J_L s^2 \Theta_L - M_L \tag{9.35}$$

Note that the compliance $1/K_D$ appears in only one term of one equation. To find the roots of the system, these equations are to be combined to form one equation in one variable. For the purpose of this argument, the exact form of the equation need not be determined. The position of the variable $1/K_D$, however, is kept clear-cut by the following sequence of manipulations. Multiply (9.32) by $K/(1 + T_1 s)$ and subtract from (9.33) to eliminate the variable Θ_e from (9.33). Similarly, use the new form of (9.33) to eliminate the variable Θ_M from (9.34). Finally use (9.35) to eliminate Θ_L from (9.34). The final form of (9.34) is in terms of M_L as the only variable. The coefficient of M_L will be the sum of many terms, some of which contain $1/K_D$. The equation can be expressed in the form

$$0 = F_1(s) \pm \frac{1}{K_D} F_2(s) \tag{9.36}$$

This form is possible because $1/K_D$ cannot appear as a squared term, as shown by the sequence of elimination of variables described above. The roots of this characteristic equation for the special case of $1/K_D$ being zero are the zeros of the function $F_1(s)$. The denominators of any terms which might appear in the final form of (9.34) are cleared; the functions $F_1(s)$ and $F_2(s)$ in (9.36) have no denominators. For $1/K_D$ infinite, the only values of s which can be roots are values that make $F_2(s)$ equal to zero. Denoting the first set of roots as $p_1, p_2, \ldots,$ (9.36) can be written in the form

$$\left(1 - \frac{s}{p_1}\right)\left(1 - \frac{s}{p_2}\right) \cdots \pm \frac{1}{K_D}\left(1 - \frac{s}{q_1}\right)\left(1 - \frac{s}{q_2}\right) \cdots = 0 \tag{9.37}$$

The sum of these two terms can be made zero for intermediate values of $1/K_D$ by making the ratio of the two terms equal to ±1.

$$\frac{(1/K_D)(1 - s/q_1)(1 - s/q_2) \cdots}{(1 - s/p_1)(1 - s/p_2) \cdots} = \pm 1 \tag{9.38}$$

The sign of ±1 can usually be easily determined in any particular system by some physical condition which is satisfied for one choice of sign but cannot be satisfied by the other. This line of attack thus produces the same conclusion as that involving block diagrams.

The use of determinants to solve a set of simultaneous equations offers many advantages in organizing a complex system and will be the next topic of study.

CHAPTER 10

USE OF DETERMINANTS

A block diagram of a system emphasizes the cause-and-effect sequence of signals in the system. Thus a block diagram can be built up step by step without ever thinking of the complete set of simultaneous equations involved. The circuit method of the last chapter treats the entire system from the point of view of one element, and system behavior is then described completely by the roots of the system for the extreme cases when that one variable is zero and infinite. Determinants, on the other hand, are a shorthand method of dealing with simultaneous equations. The manipulations with them correspond to setting up new equations, either by adding two equations or by changing variables.

Each of these methods is just a different way of solving the same problem. The operations in one method would therefore be expected to be similar to or the same as those of another, as will be shown to be the case. The value of one method compared to another depends upon the clarity with which it presents the system and the pattern of solution which it suggests. The expansion of a determinant, for example, may be done in any one of several ways. Considerable attention will be paid to obtaining the simplest sequence.

10.1. Correlation with Block Diagrams

A block diagram emphasizes cause and effect; a determinant normally expresses a set of simultaneous equations. By proper arrangement of the determinant, however, the corresponding block diagram is immediately evident.

The purpose of this article is to demonstrate that a determinant can be interpreted as a block diagram. Consider the autopilot system shown in Fig. 8.26 for the special case of K_γ and K_v equal to 0. Lower-case letters λ_o, γ, α, and β are used in the block diagram because time-varying signals are here pictured. For equations, however, the corresponding capital letters will be used: Λ_o, Γ, A, and B. The change corresponds to considering signals of the form e^{st} throughout, such as $\alpha = \mathrm{A}e^{st}$. Starting with λ_o as the independent variable in this case, the equations at each succeeding junction from left to right are then

$$(C_T\Lambda_o - C_M\mathrm{A})\frac{1}{Js^2} = \Gamma \tag{10.1a}$$

$$\Gamma - \mathrm{B} = \mathrm{A} \tag{10.1b}$$

$$\frac{C_L}{MVs}\mathrm{A} = \mathrm{B} \tag{10.1c}$$

Note that, as one reads these equations, the sequence of cause and effect is still preserved, in that the variable determined by an equation is stated last. The pattern of equations for use in determinants, however, requires that all terms involving a given variable be listed in one column, as shown by

$$-Js^2\Gamma - C_M\mathrm{A} \qquad\qquad = -C_T\Lambda_o \tag{10.2a}$$

$$\Gamma \quad - \mathrm{A} \quad\quad - \mathrm{B} = 0 \tag{10.2b}$$

$$C_L\mathrm{A} - MVs\mathrm{B} = 0 \tag{10.2c}$$

The independent variable for this case is listed in the right-hand column; the other three variables are all dependent. Note the pattern that the variable Γ determined by the first equation is listed in the first column; the variable A determined by the second equation is listed in the second column, etc. The variable determined by each equation is thus located along the so-called "principal diagonal" from upper left to lower right of the array. The determinant itself is simply the array of coefficients which remains after eliminating the variables, although the variables are identified at the top of the columns in Fig. 10.1.

Γ	A	B		Λ_o
$-Js^2$	$-C_M$	0		$-C_T$
+1	−1	−1		0
0	C_L	$-MVs$		0

FIG. 10.1. Array of coefficients of simultaneous equations of the pitch motion of plane.

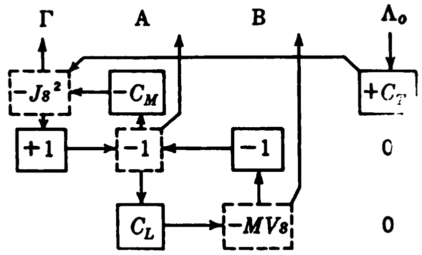


FIG. 10.2. Modification of Fig. 10.1 to indicate equivalent block diagram.

A block diagram can be regained from this figure by treating each coefficient as a transfer function, as shown in Fig. 10.2. The independent variable Λ_o multiplied by the transfer function C_T acts to produce a change in the variable Γ. Assuming that A is known, then $-C_M\mathrm{A}$ is the other torque acting to change Γ. The top row of Fig. 10.2 represents Eq. (10.2*a*).

To determine the variable Γ, the sum of these torques must be divided by Js^2.

$$(-C_M\mathrm{A} + C_T\Lambda_o)\frac{-1}{-Js^2} = \Gamma \tag{10.3}$$

The significance of the dashed block in Fig. 10.2, therefore, is that the input signals must be multiplied by the negative reciprocal of the function in this block to obtain the variable in that column. The sign of the coefficient of the input variable Λ_o is changed because this term is brought back to the left side of the equation. The same reasoning is applied to the other rows, thus setting up the pattern of signal flow as shown. This symmetrical arrangement of transfer functions is such that the arrows need not be drawn; the position of the transfer function alone specifies its meaning. The control equation which determines λ_o must also follow this pattern (Fig. 10.3). The block diagram in Fig. 10.3 now is identical to that in Fig. 8.26 except for the use of the symbols F_γ for $K_\gamma G_\gamma$, F_β for $K_y G_y V/s$, F_λ for the inverse transfer function between λ_i and λ_o, and λ for the input at the extreme left of Fig. 8.26. One advantage of the use of determinants is that the array is set by a given choice of columns for the variables whereas a block diagram might be drawn in many ways.

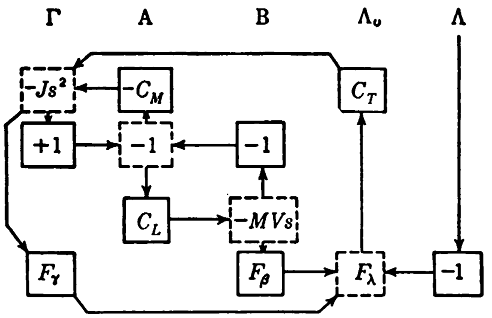


FIG. 10.3. Modification of Fig. 10.2 to include autopilot control functions.

10.2. Meaning of Determinant Manipulations

A determinant is a shorthand form of expressing simultaneous equations in which each row represents an equation, each column a variable, and each element of the array a coefficient. Addition of rows involves addition of equations; additions of columns involves changing variables. Manipulations which produce a zero element eliminate the effect of the variable in that column from the equation in that row.

The general form of the pitch autopilot system is shown in Fig. 8.32, in which a function such as a_{12} represents the transfer function to θ_1 from θ_2 along the given path. The elimination of the variable θ_2 will be achieved by determinant manipulations to provide a comparison with the block-diagram method in Art. 8.10. The equations at each junction point for symbols of the form

$$\theta_i = \Theta_i e^{st} \tag{10.4}$$

are

$$-\Theta_x = -\Theta_1 + a_{12}\Theta_2 + a_{13}\Theta_3 \tag{10.5a}$$
$$0 = a_{21}\Theta_1 - \Theta_2 + a_{23}\Theta_3 \tag{10.5b}$$
$$0 = a_{31}\Theta_1 + a_{32}\Theta_2 - \Theta_3 \tag{10.5c}$$

Note that the equations are written so that each variable appears in a separate column; omitting the variables leaves the array of coefficients shown in Fig. 10.4.

Each row represents an equation with the variables understood to be those given at the tops of the columns in Fig. 10.4. A new equation in which the coefficient of Θ_2 is zero may be obtained by multiplying each coefficient in the second equation by a_{12} and adding them to corresponding coefficients in the first equation. Similarly, multiplying each coefficient in the second row by $+a_{32}$ and adding to the third row eliminates the coefficient of Θ_2 in the third equation. The determinant array corresponding to the new equations is shown in Fig. 10.5. The variables Θ_1 and Θ_3 can now be determined from the first and third equations with the second equation merely determining Θ_2 from Θ_1 and Θ_3.

$$\begin{array}{ccc} \Theta_1 & \Theta_2 & \Theta_3 \\ -1 & a_{12} & a_{13} \\ a_{21} & -1 & a_{23} \\ a_{31} & a_{32} & -1 \end{array}$$

FIG. 10.4. The array of coefficients for simultaneous equations corresponding to the block diagram in Fig. 8.32.

$$\begin{array}{ccc} \Theta_1 & \Theta_2 & \Theta_3 \\ -1+a_{12}a_{21} & 0 & a_{13}+a_{12}a_{23} \\ a_{21} & -1 & a_{23} \\ a_{31}+a_{32}a_{21} & 0 & -1+a_{32}a_{23} \end{array}$$

FIG. 10.5. The array of coefficients for simultaneous equations corresponding to the block diagram in Fig. 8.34.

Another procedure to eliminate Θ_2 involves addition of columns of coefficients. Thus if the second column of Fig. 10.4 is multiplied by a_{21} and added to corresponding elements of the first column, the element in the position of a_{21} will now be zero. The equations will remain valid, however, only if a variable is changed. Specifically the first equation, represented by the first row in Fig. 10.4, would be greater by the quantity $a_{21}a_{12}\Theta_1$ arising in the first column but would be less by exactly the same amount since the variable in the second column was changed from Θ_2 to $\Theta_2 - a_{21}\Theta_1$. Repeating the procedure by multiplying elements in the second column of Fig. 10.4 by a_{23} and adding them to elements in the third column gives the determinant shown in Fig. 10.6. The variable in the second column is now zero. The coefficients in the second column

$$\begin{array}{ccc} \Theta_1 & \begin{array}{c}\Theta_2 \\ -a_{21}\Theta_1 \\ -a_{23}\Theta_3\end{array} & \Theta_3 \\ -1+a_{12}a_{21} & a_{12} & a_{13}+a_{12}a_{23} \\ 0 & -1 & 0 \\ a_{31}+a_{32}a_{21} & a_{32} & -1+a_{32}a_{23} \end{array}$$

FIG. 10.6. Elimination of one equation by change in variable from Fig. 10.4.

of Fig. 10.6 can thus be eliminated from the first and third rows without changing these equations.

10.3. Development of Cramer's Rule

> *The transfer function from an input to an output of a system can be expressed as a ratio of determinants. A determinant is a square array of elements whose value is the sum of all possible products involving one element from each row and column with proper choice of sign.*

The manipulation of equations in the previous article demonstrates that a set of equations can be reduced to a single equation in a single dependent variable; the functions obtained are the sums of various products of elements in the array of coefficients. The purpose of this article is to investigate the form which these functions must have, before considering a proof.

In solving for the transfer function from an input to an output, the values of s which make the denominator zero correspond to output transients which can exist when the input is zero. But if the input is zero, its column of coefficients can be neglected. The remaining array of coefficients is called the characteristic determinant of the system and is independent of the output being determined. The values of s which make the numerator function zero, however, correspond to inputs of the form e^{st} which can exist when the output is zero. In this case, the column of coefficients for the output variable will have no effect upon the solution. The problem of finding the zeros of the transfer function is thus separated from the problem of finding the poles. In each case a set of three equations in three variables must be solved by reducing each set to a single equation in a single variable. Any of the manipulations of the previous article may be used because the only object at this point is to find the roots of the equations.

A single equation may be obtained in a single variable by following a consistent pattern for obtaining zero elements. If the procedure of multiplying rows by an element and adding rows, which was used in obtaining Fig. 10.5, is continued, no product will arise which has two elements of the same row involved. Similarly, if the procedure of multiplying and adding columns, which was used in obtaining Fig. 10.6, is continued, no product will arise which contains two elements in the same column. Either procedure will produce the same final function; therefore, no product in this function can contain two elements in the same row or column.

The product of elements such that no row or column is duplicated suggests a concept of feedback paths, as illustrated in Fig. 10.7. The elements in the principal diagonal are all made -1 for convenience. An element such as a_{21} is then the transfer function from θ_1 to θ_2. The closed path of a signal then involves a transfer function from each row or column only once, as shown by the arrows. The transfer function around the loop is $a_{21}a_{32}a_{13}$ and is positive. For a third-order determinant of the form of Fig. 10.7, in which the only minus sign occurs along the principal diagonal, all the loop transfer functions are positive. The signs of the terms thus obtained are consistent with those which will be given for determinants in the next article.

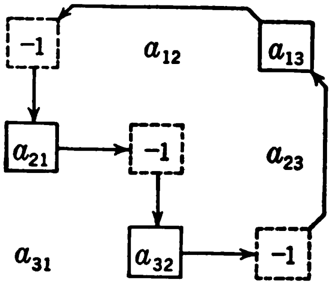


FIG. 10.7. A term in a determinant expansion identified as a transfer function around a loop.

The above correlations between block diagrams and determinants are given to show the similarity between the two. The block-diagram concept is useful in writing the equations on the basis of cause and effect. A determinant can then be used simply as a logical method of arrangement of a block diagram.

The transfer function between any input and any output is conveniently found by use of Cramer's rule. This rule is stated and shown to be valid in a typical example in the next article.

10.4. Statement and Proof of Cramer's Rule

Cramer's rule is a procedure for expressing the solution of a set of simultaneous equations in terms of the coefficients of these equations. The procedure is illustrated for a system of three equations. The proof consists of showing that the result given by the rule does satisfy the equations.

Three equations in general form are

$$a_{10}x_0 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \tag{10.6a}$$
$$a_{20}x_0 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \tag{10.6b}$$
$$a_{30}x_0 = a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \tag{10.6c}$$

The independent variable is x_0, in that it would be an input to a control system; the variables x_1, x_2, and x_3 are dependent, in that they are outputs determined by the input.

Cramer's rule gives a single equation in a single dependent variable by expressing the function in determinant form:

$$\begin{vmatrix} a_{10} & a_{12} & a_{13} \\ a_{20} & a_{22} & a_{23} \\ a_{30} & a_{32} & a_{33} \end{vmatrix} x_0 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} x_1 \tag{10.7}$$

Note that the determinant associated with the dependent variable is the array of coefficients of the set of simultaneous equations. The determinant associated with the independent variable is the same array with the column of coefficients of the dependent variable x_1 replaced by those of the independent variable x_0.

A determinant is a sum of products, in which each product, such as $a_{11}a_{23}a_{32}$, contains one and only one element from each row and column. A convenient method for obtaining all such products and giving them the proper signs is expansion by minors. A typical term in such an expansion is

$$a_{11}d_{11} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \tag{10.8}$$

The determinant d_{11} in (10.8) is called a "minor." Note that it is the same as the previous determinant except that the row and column of the factor a_{11} are eliminated. By selecting as outside factors all the elements of a particular row or column, the sum of the terms such as (10.8) will include all possible terms in the over-all determinant.

$$D_1 = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \tag{10.9}$$

Note that the signs of the terms alternate, starting with plus for the term involving the element in the first row and column. This procedure of expansion by minors can be continued until all possible products are obtained. For the given set of three equations, this is not necessary unless the expansion of a second-order determinant is considered to be expansion by minors.

$$\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = a_{22}|a_{33}| - a_{23}|a_{32}| \tag{10.10}$$

The total expansion of a determinant is then

$$D_1 = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22}) \tag{10.11}$$

Note that all possible products involving one and only one element from each row and column have been obtained by this procedure of expansion by minors.

A proof of Cramer's rule is to show that all the original simultaneous equations are satisfied. In this proof use will be made of the concept of minors to abbreviate the algebra.

First consider the special case in which the input variable acts in only the first equation; *i.e.*, a_{20} and a_{30} are zero. This permits the solutions for the dependent variables to be expressed as

$$x_1 = \frac{a_{10}d_{11}}{D_1}x_0 \qquad x_2 = \frac{-a_{10}d_{12}}{D_1}x_0 \qquad x_3 = \frac{a_{10}d_{13}}{D_1}x_0 \tag{10.12}$$

Substituting these solutions into the first equation (10.6*a*) gives

$$a_{10}x_0 = \frac{a_{10}}{D_1}(a_{11}d_{11} - a_{12}d_{12} + a_{13}d_{13})x_0 \tag{10.13}$$

But the determinant D_1 can be expressed as the sum of the terms in parentheses; thus (10.13) is satisfied.

Substitution of the values of the variables from (10.12) into the second equation (10.6*b*) gives

$$a_{20}x_0 = \frac{a_{10}}{D_1}x_0(a_{21}d_{11} - a_{22}d_{12} + a_{23}d_{13}) \tag{10.14}$$

The sum of the terms in parentheses in this case could be expressed as a single third-order determinant in which the elements such as a_{21} are put into the first row, as shown in (10.15):

$$a_{21}d_{11} - a_{22}d_{12} + a_{23}d_{13} = \begin{vmatrix} a_{21} & a_{22} & a_{23} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \tag{10.15}$$

Expanding this determinant by minors about the first row gives the sum of the terms shown on the left. But the value of the determinant is not changed by subtracting the second row from the first row, giving

$$\begin{vmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = 0 \tag{10.16}$$

Note that the new first row has all elements zero, and therefore expansion by minors about this row would give zero as the value of the over-all determinant; hence the right side of (10.14) is zero. Thus Eq. 10.6*b* is satisfied by the solutions obtained by Cramer's rule, since a_{20} was taken as zero. The same procedure will show that Eq. (10.6*c*) is also satisfied in the special case under discussion.

Consider now the case in which a_{20} has a definite value but a_{10} is now zero. Repetition of the entire argument will show that all equations are again satisfied. Similarly, if a_{30} has a definite value, with a_{10} and a_{20} both zero, all equations are satisfied. If a_{10}, a_{20}, and a_{30} are all present, the solutions for x_1, x_2, and x_3 will be the sum of the solutions for each term individually. By superposition, the complete form of the solutions for x_1, x_2, and x_3 will satisfy all three equations.

The argument used here for an example of a system with three simultaneous equations apply equally well to a system with any number of equations. Thus when solutions are expressed in the form using minors, as in (10.12), no restriction is made regarding the order of the minor.

10.5. Pitch Autopilot System

The solution of the pitch autopilot system achieved in Chap. 8 by block-diagram manipulation is easily followed in determinant form. The addition of a lift force from the elevator involves adding one more term to the determinant.

The pitch autopilot system shown in block-diagram form in Fig. 8.26 is shown in determinant form in Fig. 10.8.

The column at the left gives the various inputs which might act on the system. The first row corresponds to the moment equation, the third row to the lift equation, and the fourth row to the control equation. The first simplifications of the system are carried out in terms of these equations; determinant-expansion rules are not involved until after application of Cramer's rule. Thus A is eliminated by multiplying the second row of coefficients by $-C_M$ and adding to the first row, and multiplying the second row by C_L and adding it to the third row, to give Fig. 10.9.

$$\begin{array}{ccccc}
 & \Gamma & A & B & \Lambda_o \\
 & -Js^2 & -C_M & 0 & C_T \\
0 & +1 & -1 & -1 & 0 \\
F_i & 0 & C_L & -MVs & 0 \\
\Lambda & F_\gamma & 0 & F_\beta & F_\lambda
\end{array}$$

FIG. 10.8. The pitch autopilot system of Fig. 8.26 shown in determinant form.

Divide the first row of Fig. 10.9 by C_M and the second row by C_L to give Fig. 10.10. The new terms which are used are identified by their positions; thus $\omega_1 = \sqrt{C_M/J}$, $R_T = C_T/C_M$, and $T_\beta = MV/C_L$.

$$\begin{array}{cccc}
 & \Gamma & B & \Lambda_o \\
 & -(Js^2+C_M) & C_M & C_T \\
F_i & C_L & -(C_L+MVs) & 0 \\
\Lambda & F_\gamma & F_\beta & F_\lambda
\end{array}$$

FIG. 10.9. Elimination of the variable A by addition of equations in Fig. 10.8.

$$\begin{array}{cccc}
 & \Gamma & B & \Lambda_o \\
 & -(1-s/j\omega_1)(1+s/j\omega_1) & 1 & R_T \\
F_i/C_L & 1 & -(1+T_\beta s) & 0 \\
\Lambda & F_\gamma & F_\beta & F_\lambda
\end{array}$$

FIG. 10.10. Modification of Fig. 10.9 using dimensionless ratios.

The locked-control case has F_γ and F_β zero; the value of the characteristic determinant is then simply

$$\Delta_1 = F_\lambda \left[\left(1 - \frac{s}{j\omega_1}\right)\left(1 + \frac{s}{j\omega_1}\right)(1 + T_\beta s) - 1 \right] \qquad (10.17)$$

The zeros of the function in brackets are determined in the root-locus plot in Fig. 8.28 and are denoted as s_1 and s_2. The full value of the determinant in Fig. 10.10, as expressed by expansion by minors about the third row, is

$$\Delta_2 = F_\lambda\left(1 - \frac{s}{s_1}\right)\left(1 - \frac{s}{s_2}\right)s + F_\gamma R_T(1 + T_\beta s) + F_\beta R_T \tag{10.18}$$

The solution in Art. 8.10 involves combining the two control functions into a single function, as expressed in (8.26). The zeros of this function are q_1 and q_2, and a lead network is then inserted in series of the form $K_C(1 + Ts)/(1 + \alpha Ts)$. The inverse transfer function between λ_i and λ_o is $(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)$, making the determinant now take the form

$$\Delta_2 = \left(1 - \frac{s}{r_1}\right)\left(1 - \frac{s}{r_2}\right)\left(1 - \frac{s}{r_3}\right)\left(1 - \frac{s}{s_1}\right)\left(1 - \frac{s}{s_2}\right)s + \frac{R_T K_v V(1 - s/q_1)(1 - s/q_2)}{s}\,\frac{K_C(1 + Ts)}{1 + \alpha Ts} \tag{10.19}$$

The roots of this equation are the values of s which make the ratio of the two terms equal to -1, the latter equality being identical to (8.27). The roots are determined in Fig. 8.30. If the transfer function between Λ and Γ is desired, Cramer's rule is used for the determinant in Fig. 10.10. The numerator determinant then has only one term, $R_T(1 + T_\beta s)$, making $\Gamma/\Lambda = R_T(1 + T_\beta s)/\Delta_2$. Cramer's rule could also be invoked to find the transfer function from any other input to any other output. Another advantage of using determinants is the procedure for adding another coupling in the system, as discussed next.

The effect of the lift force of the elevator is neglected in Art. 8.10. This force can be included in the equations represented in Fig. 10.8 by changing 0 to $-C_R$ in the third row, Λ_o column. In Fig. 10.10 the term would appear as $-R_R = -C_R/C_L$ in the second row, Λ_o column. The previous solution for Δ_2 need not be repeated; any expansion of the characteristic determinant is simply greater by R_R times its minor:

$$\Delta_3 = R_R\left[-\left(1 - \frac{s}{j\omega_1}\right)\left(1 + \frac{s}{j\omega_1}\right)F_\beta - F_\gamma\right] \tag{10.20}$$

The values of F_β and F_γ used previously should again be inserted and the zeros of (10.20) found. The characteristic equation for the new system is then

$$\Delta_4 = \Delta_2 + \Delta_3 \tag{10.21}$$

This equation is then solved by a root-locus plot in which R_R is the variable. The numerator determinant of any transfer function previously determined by Cramer's rule should also be increased by R_R times its minor in that determinant. Thus Λ_o/Λ would be

$$\frac{\Gamma}{\Lambda} = \frac{-R_R + R_T(1 + T_\beta s)}{\Delta_2 + \Delta_3} \tag{10.22}$$

10.6. Roll-Yaw Coupling

> *The yaw system is described by three simultaneous equations, roll by two. Each of these systems, treated independently, can be solved readily. Some aerodynamic effects couple the two systems together, but a new solution for the complete system can be built from the individual solutions.*

The motion in yaw is described by a moment equation, a side-force equation, and a rudder-control equation. The variables are the angles shown in Fig. 10.11*a*; the equations are of the same form as those of pitch. The roll system is described by a moment equation and by an aileron-control equation; these angles are shown in Fig. 10.11*b*. Many intercoupling effects are present however. When the plane has a roll rate, for example, the angle of attack on the rising wing is decreased, which decreases the drag on that wing. The opposite effect occurs on the lowering wing, and thus a net yawing moment is created. Coupling from yaw into roll occurs with the plane yawed, because the force on the vertical tail surfaces acts above the roll axis of the plane and thus introduces a rolling moment.

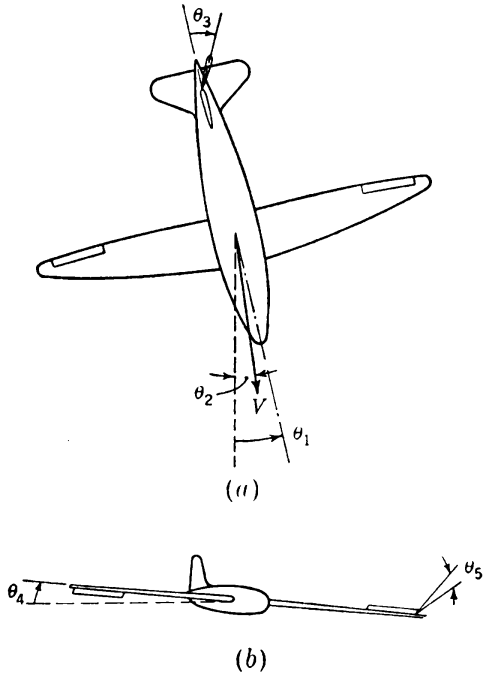


FIG. 10.11. The airplane angles involved in roll-yaw cross coupling.

The purpose of this article is to show how these intercoupling effects can be introduced, drawing upon the previous solutions for roll and yaw as independent systems. The symbols used in the determinant in Fig. 10.12 are therefore selected with subscripts to identify the row and column in which the symbols appears. The symbol C denotes a given function, and F an adjustable function. The element C_{14} arises from the yaw torque produced by the roll rate. Similarly C_{41} arises from the rolling moment produced by the yaw angle.

The dashed lines separate the yaw and the roll systems. The determinant in the upper left of Fig. 10.12 is the characteristic determinant of the yaw system, and its zeros are presumed to be known from a study of that system. Similarly the determinant in the lower right is that for the roll system, and its zeros are also presumed to be known. These determinants can be used in the solution for the complete determinant by the following procedure: Assume that Θ_1 is known; therefore the input to the roll system is $C_{41}\Theta_1$. Solve for Θ_4 using Cramer's rule.

$$\Theta_4 = \frac{\Theta_1}{\Delta_{\text{roll}}} \begin{vmatrix} -C_{41} & C_{45} \\ - & C_{55} \end{vmatrix} = \frac{\Theta_1 \Delta_{14}}{\Delta_{\text{roll}}} \tag{10.23}$$

The change in sign of C_{41} in using Cramer's rule arises because the rule is based upon the column of coefficients being on the other side of the equation rather than on the same side.

$$\begin{array}{ccc|cc l}
\theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & \\
C_{11} & C_{12} & C_{13} & C_{14} & & \text{Yaw moments} \\
C_{21} & C_{22} & & C_{24} & & \text{Side forces} \\
 & F_{32} & C_{33} & & & \text{Rudder controls} \\
\hdashline
C_{41} & & & C_{44} & C_{45} & \text{Roll moments} \\
 & & & C_{54} & C_{55} & \text{Aileron controls}
\end{array}$$

FIG. 10.12. The determinant of the roll-yaw system isolating the individual determinants.

With Θ_4 known, the inputs to the yaw system are known and Θ_1 can thus be found by use of Cramer's rule.

$$\Theta_1 = \frac{\Theta_4}{\Delta_{\text{yaw}}} \begin{vmatrix} -C_{14} & C_{12} & C_{13} \\ -C_{24} & C_{22} & C_{23} \\ & C_{32} & C_{33} \end{vmatrix} = \frac{\Theta_4 \Delta_{41}}{\Delta_{\text{yaw}}} \tag{10.24}$$

The transfer function around the complete loop obtained from (10.23) and (10.24) is equated to +1 to obtain the characteristic equation.

$$\frac{\Theta_1}{\Theta_4}(s)\,\frac{\Theta_4}{\Theta_1}(s) = \frac{\Delta_{14}}{\Delta_{\text{roll}}}\,\frac{\Delta_{41}}{\Delta_{\text{yaw}}} = +1 \tag{10.25}$$

The complete determinant would then be:

$$\Delta = \Delta_{\text{roll}}\Delta_{\text{yaw}} - \Delta_{14}\Delta_{41} \tag{10.26}$$

The complete determinant in Fig. 10.12 could have been expanded by any of the usual methods, and the final solution for the roots of the characteristic equation would, of course, be the same for the same values of the constants. The indicated procedure preserves the picture of cause and effect, however, and makes use of the individual determinants for the roll and yaw systems already solved.

The coupling loop could be broken by adding a function F_{51} such that the numerator of the transfer function from Θ_1 to Θ_4 is zero.

$$\Theta_4 = \frac{\Theta_1}{\Delta_{\text{roll}}} \begin{vmatrix} -C_{41} & C_{45} \\ -F_{51} & C_{55} \end{vmatrix} = 0 \tag{10.27}$$

Solving for F_{51} gives:

$$F_{51} = \frac{C_{41}C_{55}}{C_{45}} \tag{10.28}$$

10.7. Effect of Small Additional Couplings

> *The roll-yaw system has many small coupling effects for which a complete solution is not justified. Two simplifications may be exploited: neglecting all terms involving the product of two small numbers, and solving for the shift in the critical roots only.*

The roll-yaw system can be solved by the preceding method very conveniently when the number of coupling coefficients is small. Many other coupling effects are present, however, as indicated by the following examples: The coefficient C_{15}, for example, would introduce a yaw moment due to an aileron deflection. The coefficient C_{42} introduces the decrease in rolling moment due to change in angle of attack in yaw with sidewise velocity. The coefficient C_{23} introduces the side force due to rudder deflection. Any one of these coefficients could be added to the determinant array shown in Fig. 10.12 with only one or two terms added to the expanded form of the determinant. If all such terms are considered simultaneously, however, the minor associated with each coefficient will in turn require several plots to obtain the factors. The additional terms introduced by the latter procedure are those involving products of small terms such as $C_{15}C_{42}$. The consideration of these coefficients one at a time in the basic determinant is a convenient way to obtain the major effect of a small term.

Another simplification that may be exploited is that the basic determinant frequently has only one pair of roots, say r_1 and r_2, near the $j\omega$ axis. The substitution of this specific value for $s = r_1$ in the determinant makes each element have a specific value. The minor associated with a new coupling term can then be evaluated by the usual numerical determinant methods rather than by separate root-locus plots. The minor need be evaluated only to perhaps 20 per cent accuracy, a fact which suggests a rapid procedure using vector diagrams. Each element in the minor is shown as a vector. Multiplying a row by a vector changes all the amplitudes by the same factor and rotates all vectors through the same angle. Adding one row to another can produce zero for one element in the new row, with the other elements determined by direct vector addition. A pattern of manipulation can be developed to make all elements above the principal diagonal equal to zero. The value of the minor, denoted as $A_2e^{j\phi}$, is then simply the product of elements along the principal diagonal. The original determinant before the addition of the coupling term C_2 is denoted as Δ_1 with known zeros r_1, r_2, r_3, etc. The complete determinant in the region of $s = r_1$ is then

$$\Delta = \Delta_1 + \Delta_2 = C_1\left(1 - \frac{s}{r_1}\right)\left(1 - \frac{s}{r_2}\right) \cdots + C_2A_2e^{j\phi} \qquad (10.29)$$

Expressing $1 - s/r_1$ as $(s - r_1)/-r_1$, the change $s - r_1$ in the position of the root r_1 can be found directly from (10.29):

$$s - r_1 = \frac{-C_2A_2e^{j\phi}}{[C_1(-1/r_1)(1 - s/r_2) \cdots]_{s=r_1}} \qquad (10.30)$$

Very often the change $s - r_1$ thus found will be small, and therefore a more complete solution is not justified.

10.8. The General Problem

> *For an nth-order determinant, the number of terms obtained by direct expansion is n!; if each element is quadratic, the degree of a polynomial is only 2n. The general methods of polynomial solution should therefore be used; a root-locus plot is useful only for the case of two terms each of which is already factored.*

The root-locus plot was introduced in Chap. 7 for a single-loop control system; the plot is ideal for that problem, because the open-loop transfer function is known in factored form and loop gain is the main variable of interest. The applications of the plots to multiloop systems in Chap. 8, to circuits in Chap. 9, and to determinants in this chapter complete this study. This final topic logically demonstrates that the other methods are preferable for many problems. Thus an nth-order determinant would have $n!$ terms, which would require $n! - 1$ plots. On the other hand an expansion into a polynomial would make the degree of the highest term $2n$ if each element were a quadratic. Substitution of numerical values makes the expansion into a polynomial comparatively fast; the effect of any one parameter can be seen by similarly computing the minor associated with that parameter. For finding the zeros of an mth-degree polynomial, the method described in Art. 8.5 would require $m - 1$ plots. The total number of plots can be reduced by about two by factoring, as in

$$\begin{aligned} &a_5s^5 + a_4s^4 + a_3s^3 + a_2s^2 + a_1s + a_0 \\ &\qquad = s^4(a_5s + a_4) + s^2(a_3s + a_2) + (a_1s + a_0) \qquad (10.31) \end{aligned}$$

If quadratic functions are factored separately, then the number of plots for an mth-degree polynomial can be reduced to about $m/3$ by expanding, as in

$$\begin{aligned} &a_5s^5 + a_4s^4 + a_3s^3 + a_2s^2 + a_1s + a_0 \\ &\qquad = s^3(a_5s^2 + a_4s + a_3) + (a_2s^2 + a_1s + a_0) \qquad (10.32) \end{aligned}$$

A more rapid method of finding the approximate values of the roots, however, is to plot each term of the polynomial versus the magnitude of s on a log-log scale, as in Fig. 10.13. Each of the coefficients is plotted on the line for $s = 1$. The term $a_1 s$ is then plotted through a_1 with a slope of 1, the $a_2 s^2$ term through a_2 with a slope of 2, etc. The largest terms for each range of values of s are then evident. At point A the terms a_0 and $a_1 s$ are larger than all others and are equal to each other, making an approximate root $s = -a_0/a_1$. At point B the terms $a_1 s$ and $a_3 s^3$ are the largest and are equal to each other; the magnitude of s can be read on the horizontal scale or computed as $\sqrt{a_1/a_3}$. If s were $j\omega$, the two terms would cancel, leaving the $a_2 s^2$ term as the largest term in the function. Actually a complex value of s can be determined from a vector picture such as Fig. 10.14.

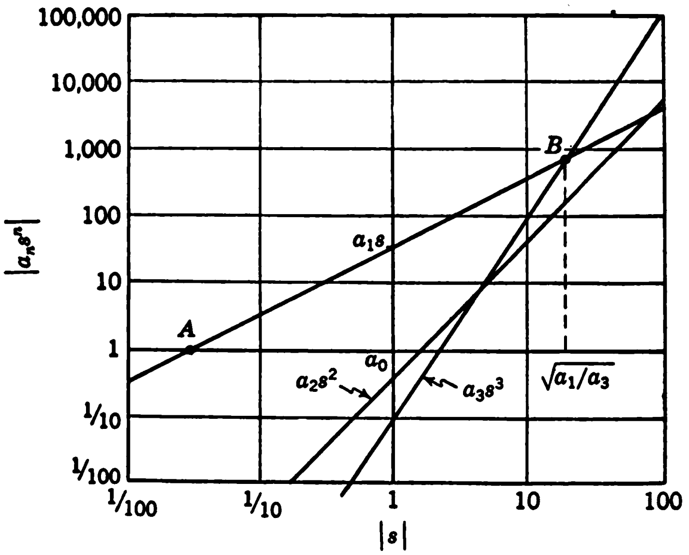


Fig. 10.13. Approximate solution for roots of a polynomial.

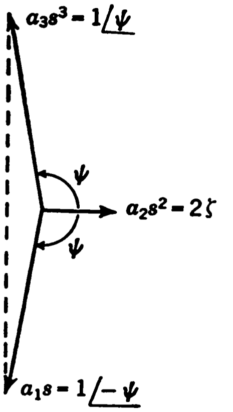


Fig. 10.14. Vector plot of terms in Fig. 10.13 for $|s| = \sqrt{a_1/a_3}$.

Considering the $a_1 s$ and $a_3 s^3$ terms as having a magnitude of unity, then the $a_2 s^2$ term must have the value 2ζ. After the approximate value of all roots are thus determined, several procedures are possible to obtain greater accuracy. At point B, for example, the remaining value of the function is a_0. The derivative of the polynomial is another polynomial, which can also be plotted on a log-log scale versus the magnitude of s. The value of the derivative for the approximate root determined for B can thus be computed by using the largest terms and vector addition if necessary. A correction to the root is then the remainder a_0 divided by the derivative.

An interesting possibility for high-order determinants is to identify the largest terms without need for expansion of the determinant. Any method which first obtains an approximate solution, with a logical procedure for improving the accuracy of this solution, has several advantages over one which requires a complete expansion first. The approximate solution is less likely to involve numerical errors, and the analysis can be terminated as soon as the required accuracy is obtained.

CHAPTER 11

TRANSFER FUNCTIONS AND TRANSIENT RESPONSE

The analysis of a system in Chaps. 7 through 10 is carried to the point of determining the zeros and poles of the transfer function of the system. The transient response is a sum of terms of the form e^{st}, one for each pole, with the amplitude given by the formula (4.21). The plots in the complex plane are an aid in evaluating these amplitudes, as illustrated in the solution of previous problems. In Art. 6.5 the amplitude is found by incremental lengths on a conformal map of the transfer function. In Art. 7.9, an amplitude is found by the ratio of vector lengths on the root-locus plot. In Art. 7.10 the effect of a zero and pole pair is evaluated. These methods are generalized in this chapter to apply to any problem in which the input is a simple function.

If the input is itself an arbitrary function of time, several possibilities are open. First, the curve can be approximated by a sum of exponential terms, so that its transform may be found, thus permitting the output to be evaluated by the same methods as above. Second, the frequency spectrum of the input and output could be used. Third, the curve might be used directly by a graphical procedure. Of these three, the first and third are considered in this chapter because the frequency-response method is well treated in other texts.

11.1. Amplitude of Transient Term from Known Zeros and Poles

> *The amplitude of a transient term for a given pole is proportional to the product of vectors from the zeros to that pole and inversely proportional to the product of vectors from all other poles to that pole.*

The net result of an analysis in which the roots of the characteristic equation are found is that the transfer function from input to output is known in factored form.

$$\frac{\Theta_o}{\Theta_i}(s) = \frac{(1 - s/q_1)(1 - s/q_2)\cdots}{(1 - s/p_1)(1 - s/p_2)\cdots} \tag{11.1}$$

If the input is some simple function, a unit step, for example, it can be considered to have been produced by pure integration of a unit impulse. The transform of the input, or the transfer function between a unit impulse and the input, is then $1/s$. The transform of the output, or the complete transfer function from a unit impulse to the output, is then

$$\Theta_o(s) = \Theta_i(s)\left[\frac{\Theta_o}{\Theta_i}(s)\right] = \frac{1}{s}\,\frac{(1 - s/q_1)(1 - s/q_2)\cdots}{(1 - s/p_1)(1 - s/p_2)\cdots} \tag{11.2}$$

The inverse transform, or the output function of time, is given in a table of Laplace transforms or obtained as described in Art. 4.6. The result can be expressed, in general, as

$$\theta_0 = A_0e^{p_0t} + A_1e^{p_1t} + A_2e^{p_2t} + \cdots \tag{11.3}$$

A term exists for each value of s which makes the denominator of (11.2) zero, and the amplitude of the ith term is given by

$$A_i = [\Theta_o(s)(s - p_i)]_{s=p_i} \tag{11.4}$$

Note that this amplitude is obtained from the transfer function by first canceling the term which would make the denominator zero $(s - p_i)$, and then substituting $s = p_i$ throughout. The amplitude A_0 is calculated from (11.2) for $s = 0$, with s omitted from the denominator, giving $A_0 = 1$.

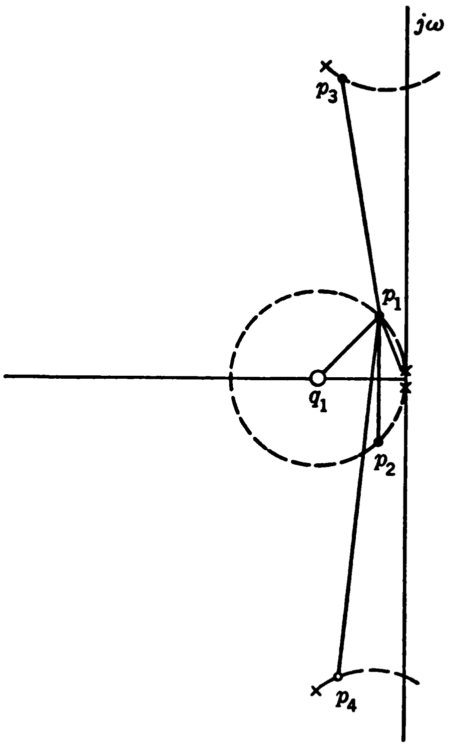


FIG. 11.1. Vectors involved in computing amplitude A_1 of transient term $A_1e^{p_1t}$.

This formula for amplitude involves vector calculations which are the same as those used in computing gain in a root-locus plot. Consider the example of the plot shown in Fig. 11.1, in which $p_1 = -0.6 + j1.5$ and $p_3 = -1.5 + j7.2$. The amplitude of the term A_1 is then

$$A_1 = \frac{(s - q_1)/(-q_1)}{(p_1)\left(\dfrac{1}{-p_1}\right)\left(\dfrac{p_1 - p_2}{-p_2}\right)\left(\dfrac{p_1 - p_3}{-p_3}\right)\left(\dfrac{p_1 - p_4}{-p_4}\right)} \tag{11.5}$$

Note that the factor p_1 is canceled by the factor $1/p_1$, which is present because the input is a unit step. The remaining factors are all ratios of lengths from other poles or zeros to the given poles compared to their lengths to the origin. Thus $p_1 - p_2$ is the vector from p_2 to p_1, and $-p_2$ is the vector from p_2 to the origin. Such a ratio is itself a vector, whose magnitude is the ratio of the magnitudes and whose angle is the difference of the angles. These values are easily estimated by eye from the diagram, as illustrated by the substitutions:

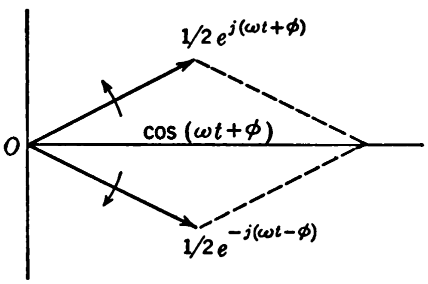
$$A_1 \cong \frac{1\underline{/45^\circ}}{(-1)(1.9\underline{/20^\circ})(1.2\underline{/10^\circ})(0.8\underline{/-10^\circ})} \cong -0.6\underline{/25^\circ} \qquad (11.6)$$

The calculation for the pole p_2 would be identical except that the phase angles would all be opposite in sign. The combination of these two terms is then simplified because they are complex conjugates. The j parts thus always cancel and the real parts add, making their sum twice the real part of either of them (Fig. 11.2).

$$A_1e^{p_1t} + A_2e^{p_2t} = 2\mathcal{R}(A_1e^{p_1t}) = 2(-0.6)e^{-0.6t}\cos(1.5t + 25^\circ) \qquad (11.7)$$

The same combination procedure can be applied to the sum of the terms involving the conjugate complex roots p_3 and p_4. Note that the root p_3, being farther from the poles p_1 and p_2 than these poles are from the origin, causes the amplitude A_3 to be small. The complete solution becomes

$$\begin{aligned}\theta_o = 1 &- 1.2e^{-0.6t}\cos(1.5t + 25^\circ) \\ &- 0.1e^{-1.5t}\cos(7.2t - 135^\circ) \qquad (11.8)\end{aligned}$$



$\frac{1}{2}e^{j(\omega t+\phi)}$
$\cos(\omega t+\phi)$
O
$\frac{1}{2}e^{-j(\omega t-\phi)}$

FIG. 11.2. Cosine wave as the sum of two oppositely rotating vectors.

An approximate criterion of the effect of a pole is that poles nearest the origin correspond to transient terms with large amplitudes. A pole near a zero involves a small transient term. Two adjacent poles produce two large transient terms which nearly cancel, as demonstrated by the case of equal roots in Art. 11.3. The method in this article requires that all the poles be known. A method which is occasionally more convenient determines the amplitude directly from the root-locus plot of the open-loop function, as discussed in the next article.

11.2. Transient from Root-locus Plot

The amplitude of a transient term can be determined directly from the root-locus plot of the open-loop function. The amplitude for a unit-step input is the ratio of the percentage shift in the root divided by the percentage change in gain which causes this shift.

The formula used in the previous section requires that all the roots be known in order to find the amplitude of a transient term for a single root. In the conformal-mapping method of finding roots, however, it is necessary and possible to find the amplitude of a term if only that particular root is known. The formula (6.9) involves the derivative of the transfer function with respect to s, which may be evaluated directly from the plot (Fig. 6.9). The same type of result can be expected for the root-locus plot. The formula in (11.4) will be evaluated for the case of $\Theta_o(s) = (\Theta_o/\Theta_i)(s)\Theta_i(s)$, in which $\Theta_i(s) = 1/s$. The evaluation of $(\Theta_i/\Theta_o)(s)$ at $s = p_1$ would give zero; therefore the derivative of the function is taken to obtain the value of the function near $s = p_1$. Then $s - p_1$ is an increment Δs, which is allowed to approach zero.

$$A_1 = \lim_{\Delta s \to 0} \frac{\Delta s/p_1}{\left[\frac{\Theta_o}{\Theta_i}(s)\right]_{s=p_1} + \left[\frac{d}{ds}\frac{\Theta_i}{\Theta_o}(s)\right]_{s=p_1} \Delta s} \tag{11.9}$$

The function $(\Theta_i/\Theta_o)(s)$ is zero at $s = p_1$; therefore, only the derivative remains in the denominator. If the feedback function is 1 and if the inverse system function is taken to be $1/KG(s)$, then

$$A_1 = \frac{1/p_1}{\left[\frac{d}{ds}\left(\frac{\Theta_i}{\Theta_o}(s)\right)\right]_{s=p_1}} = \frac{1/p_1}{\left[\frac{d}{ds}\left(1 + \frac{1}{KG(s)}\right)_{s=p_1}\right]} = \frac{1/p_1}{\frac{1}{K_1}\left[\frac{d}{ds}\frac{1}{G(s)}\right]_{s=p_1}} \tag{11.10}$$

The derivative is to be evaluated as a ratio of the increment in s and the corresponding increment in the function for the system in Fig. 5.1 with the root-locus plot given in Fig. 11.3. The increment in s is a vector along the locus as shown. The increment in the function for these two points can be found by simple but careful reasoning. The values of gain marked at the two points are the values of gain necessary to make $1/KG(s)$ equal to -1 and are therefore the values of $-1/G(s)$ at these points. The increment in $1/G(s)$ is thus $-\Delta K$, and

$$A_1 = \frac{1/p_1}{(1/K_1)\Delta[1/G(s)]/\Delta s} = \frac{(1/p_1)\,\Delta s}{(1/K_1)(-\Delta K)} = \frac{\Delta s/p_1}{-\Delta K/K_1} \tag{11.11}$$

The amplitude of a transient term is thus equal to the ratio of the percentage change in the vector s to the percentage change in the gain factor K which produces it. For the plot shown in Fig. 11.3, $\Delta K/K_1 = 1$, and the amplitude can be estimated to be

$$A_1 = \frac{0.5\underline{/-30^\circ}}{-1} = -0.5\underline{/-30^\circ} \tag{11.12}$$

Note that this value checks the value obtained in (7.45) roughly.

The system with root-locus plot in Fig. 11.1 has increments marked in Fig. 11.4; applying (11.11) gives

$$A_1 = \frac{\Delta s_1/p_1}{-\Delta K/K_1} = \frac{0.5\underline{/30^\circ}}{-1} = -0.5\underline{/30^\circ} \tag{11.13}$$

Digitized by Google Original from UNIVERSITY OF MICHIGAN

Similarly the calculation of A_3 from Fig. 11.4 gives

$$A_3 = \frac{\Delta s_3/p_3}{-\Delta K/K_3} = \frac{0.07\underline{/-120^\circ}}{-1} = -0.07\underline{/-120^\circ} \qquad (11.14)$$

Note that these values check roughly with those of (11.6) for A_1 and $A_3 = -0.05\underline{/-135^\circ}$.

One advantage of this method is that the angle of the amplitude term is known immediately from the angle of the locus. The shift in the position of the roots for various values of gain may already be known in many problems in which gain variations are unavoidable. If the shift in the roots from the poles is small compared to the distance between poles, the increment in s can be taken to be the distance from a pole to its root, thus making the ratio $\Delta K/K$ equal to 1, so that only one calculation of gain is required.

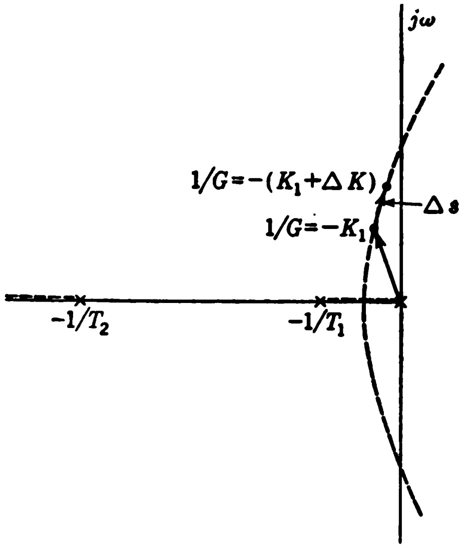


FIG. 11.3. Computation of amplitude of transient term from incremental lengths on root-locus plot.

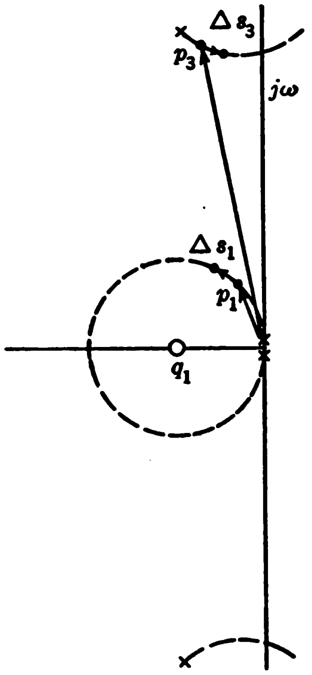


FIG. 11.4. Use of the method shown in Fig. 11.3 for the system in Fig. 11.1.

11.3. The Special Case of Equal Roots

The transient response for two equal roots can be found by considering the roots to be slightly separated and then letting the difference approach zero. The terms for the two roots prove to be $d(Ae^{st})/ds$, *in which* Ae^{st} *is the transient which would be present for a single root at that point.*

The simplest system in which two equal roots can arise is shown in Fig. 11.5. Consider the case in which the roots are nearly equal. The transient response of the system according to (11.3) is then:

$$\theta_o = \frac{r_1 r_2}{r_1 - r_2} e^{r_1 t} + \frac{r_1 r_2}{r_2 - r_1} e^{r_2 t} \tag{11.15}$$

Note that the difference $r_1 - r_2$ between the roots is very small, and therefore the amplitude of each term is very large. But the sign of the terms is different, and the transients are of approximately the same shape; therefore, the two terms nearly cancel. If $r_1 - r_2$ were allowed to approach zero in (11.15), the result would be the indeterminate result of two infinite terms which cancel perfectly.

The net result can be obtained by letting $r_1 = r_2 + dr$ and combining terms:

$$\theta_o = \frac{r_1 r_2}{dr} (e^{(r_2+dr)t} - e^{r_2 t}) = \frac{r_1 r_2}{dr} e^{r_2 t}(e^{dr\,t} - 1) \tag{11.16}$$

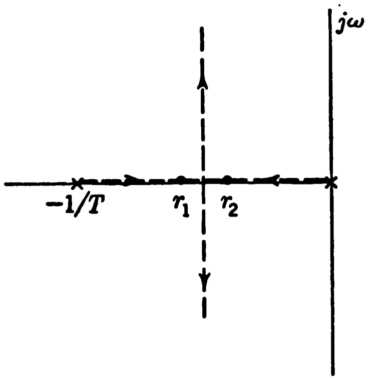


FIG. 11.5. Quadratic system having two nearly equal roots.

The function in parentheses can be expanded, giving

$$\theta_o = \frac{r_1 r_2}{dr} e^{r_2 t} \left[1 + dr\,t + \frac{(dr\,t)^2}{2} + \cdots - 1 \right] = r_1 r_2 e^{r_2 t} t \tag{11.17}$$

The term $dr\,t$ is the only significant term in the bracket, thus permitting cancellation of dr and giving the simple result shown at the right.

Consider now the general problem of finding the terms due to two equal roots when a general function $F(s)$, which may have poles, is involved.

$$\Theta_o(s) = \frac{F(s)}{(s - r_1)(s - r_2)} \tag{11.18}$$

The transient response according to (11.4) is then

$$\theta_o = \left[\frac{F(s)}{r_1 - r_2} \right]_{s=r_1} e^{r_1 t} + \left[\frac{F(s)}{r_2 - r_1} \right]_{s=r_2} e^{r_2 t} + \cdots \tag{11.19}$$

The two terms in (11.19) might also be expressed as

$$\theta_o = \frac{[F(s)e^{st}]_{s=r_1} - [F(s)e^{st}]_{s=r_2}}{r_1 - r_2} + \cdots \tag{11.20}$$

The numerator will now be recognized as being the increment in the function $F(s)e^{st}$ as s changes from r_1 to r_2. The denominator is the increment in s itself, and therefore the limit of the ratio as $r_1 - r_2$ approaches zero is simply the derivative.

$$\theta_o = \left[\frac{d}{ds}\left(F(s)e^{st}\right)\right]_{s=r_1} = \left[\frac{dF(s)}{ds}e^{st} + F(s)te^{st}\right]_{s=r_1} \tag{11.21}$$

The transient $[F(s)e^{s_1t}]_{s=r_1}$ is the transient term which would be present if only one of the equal roots were present.

11.4. Matching Initial Conditions

The operational formula for amplitudes automatically satisfies initial conditions of a system. For the special case when the input to a quadratic system is a polynomial function of time, however, the output terms can be conveniently calculated, one at a time, directly from the initial conditions.

Consider the problem of finding the output for an input of the form Vt acting on a system with an undamped sinusoidal response. The fact that the system is quadratic means that two time delays are involved. The input has an initial velocity; therefore, the output will have zero initial velocity and zero initial acceleration. The steady-state response is of the form Vt, and the transient is of the form $A \sin(\omega t + \phi)$, in which the amplitude A and phase angle ϕ are to be determined from the initial conditions. In order for the output to have no initial acceleration, the phase angle of the sine wave must be either 0 or 180°. The velocity is then maximum, and the amplitude selected is such that the velocity of the steady-state response is canceled, as shown in Fig. 11.6.

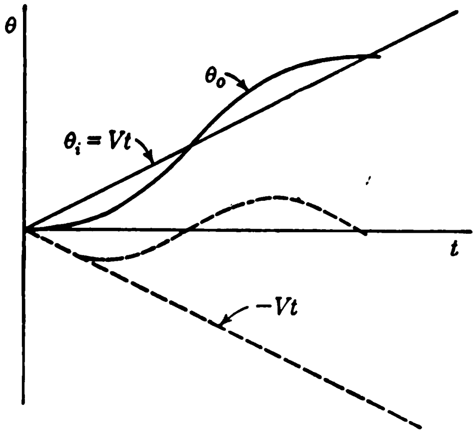


FIG. 11.6. Simple example of determining amplitudes from initial conditions.

The output will then be expressed as

$$\theta_o = Vt - \frac{V}{\omega}\sin \omega t \tag{11.22}$$

The same type of reasoning applied to a more complicated case is that of a parabolic input a_0t^2 to a system having a damped sinusoidal response. The steady-state response is a_0t^2; the transient $Ae^{\sigma t}\sin(\omega t + \phi)$ must have an amplitude and phase angle such as to cancel the acceleration in the steady-state term initially without introducing any jerk or third derivative. These conditions are conveniently determined by a vector diagram in which the projection axis for zero time is determined at the last, as shown in Fig. 11.7.

Start with the position of the amplitude vector horizontal to the right. The velocity vector is then ahead of this at an angle ψ, where $\cos \psi = \sigma/\omega_0$. The vector for acceleration is ahead of the velocity vector by ψ, and the jerk vector in turn is ahead of that. The initial value of jerk must be zero, making the real axis at zero time be the dashed line shown. The projection of the acceleration vector on this axis must cancel the acceleration of the steady-state response, which determines A from

$$A\omega_0^2 \sin \psi = 2a_0 \quad (11.23)$$

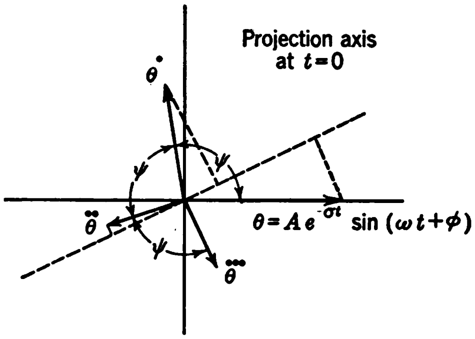


FIG. 11.7. Vector diagram to determine amplitude and phase angle of damped sinusoidal oscillation started by parabolic input.

The transient sine wave is now determined, but note that it has an initial velocity. Another transient term must arise to cancel this initial velocity; fortunately Vt will do this without upsetting the previous conditions of zero initial acceleration and jerk. The constant term C is also added to make the initial displacement zero, and again the other initial conditions are not disturbed. The total transient response is thus

$$\theta = \frac{2a_0}{\omega_0^2 \sin \psi} [e^{\sigma t} \sin(\omega t - 3\psi) - \omega_0 t \sin(-2\psi) - \sin(-3\psi)] \quad (11.24)$$

The problem could, of course, have been solved by several other methods, but the significant point is that the initial conditions can be satisfied one at a time without simultaneous equations in several variables arising.

11.5. Aids in Sketching Transients

Only two points are necessary for sketching the transient for a real root: the initial amplitude and the time constant of decay. Oscillatory roots require in addition the initial phase angle and the period of oscillation.

The formula for the transient often gives a satisfactory picture of the transient; therefore plotting is not needed. In other problems, the plot itself is necessary to determine some critical points, *e.g.*, time for the error to reach zero, peak overshoot, or time for the error to reach a given percentage of the step input. The plotting techniques are standard but are given here for completeness and as an aid in studying the reverse process of finding a transfer function from a transient curve.

Consider the example of a cubic system in which a lead network is used to increase the loop gain; the resultant root-locus plot is shown in Fig. 11.8.

Using the methods of either Art. 11.1 or Art. 11.2, the transient for a unit-step input is approximately

$$\theta_o = 1 - A_1 e^{\sigma_1 t} \cos (\omega_1 t - \phi_1) - A_3 e^{\sigma_3 t} - A_4 e^{\sigma_4 t} \tag{11.25}$$

The significant points in plotting this transient are shown in Fig. 11.9. The initial amplitudes of all terms are shown along the vertical axis. The time constant of decay of each transient, $1/|\sigma|$, is marked along the time axis. The half period of the oscillatory root is determined, and peak points and zero points are laid off from the initial phase-angle point.

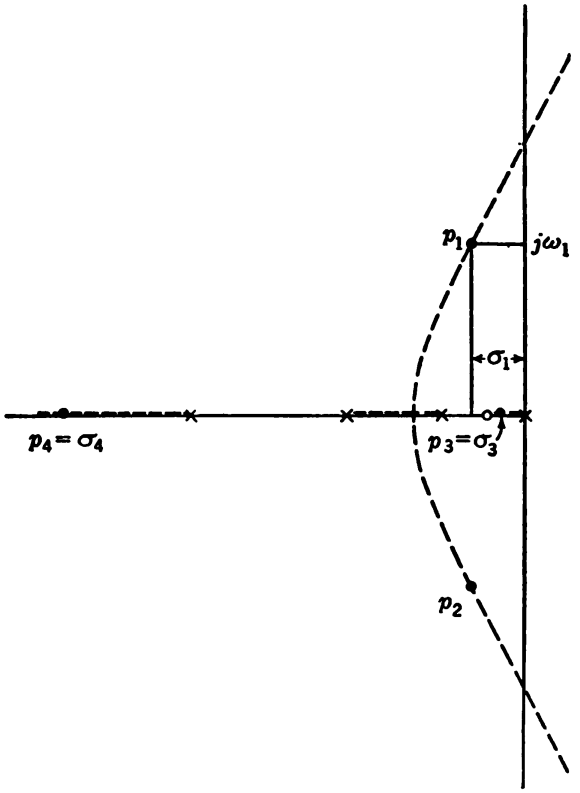


FIG. 11.8. Root-locus plot for simple system with a lead network.

In sketching the actual transients, the decay curves start tangent to the straight lines of Fig. 11.9 but then curve away, making the amplitude of the curve $1/e$ or three-eighths of the initial amplitude at the time-constant point (Fig. 11.10). The damped oscillation is sketched to be tangent to the decay envelope at the peak points of the sine wave and to cross the zero axis at the designated points.

In sketching the total transient, note that the initial amplitude is -1 and that the initial slope is zero. After the $A_4 e^{\sigma_4 t}$ term is negligible, the output curve is sketched below the oscillatory curve by the amount of the long exponential transient.

The above sketching technique gives the shape of the response and locates critical points for which specific calculations can be made. If greater accuracy is desired on the basic sketch, the use of some standard curves is helpful. A single sine wave and a family of exponential decay curves are sufficient for the given problem. The points in Fig. 11.9 are marked on a transparent sheet, which is laid over the standard curves to permit tracing the particular curve which is needed.

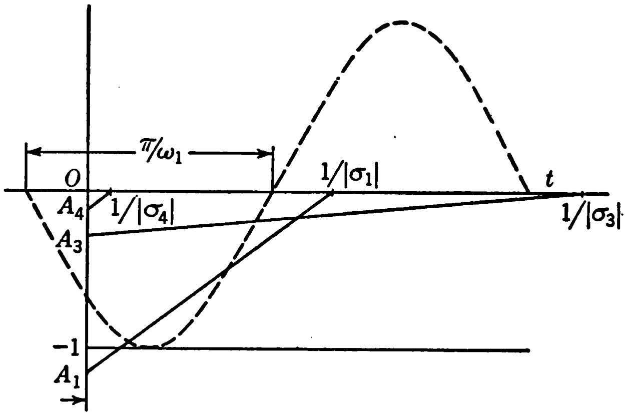


FIG. 11.9. Significant lines in constructing a plot of transient response.

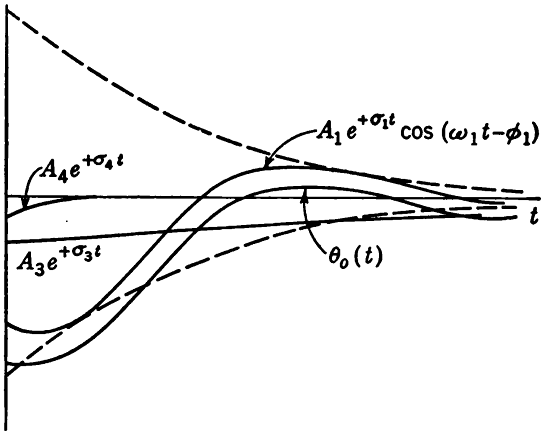


FIG. 11.10. Complete transient response sketched from Fig. 11.9.

11.6. Transfer Function from Transient Response

The dominant oscillation and decay rate of a transient can usually be identified by inspection. The corresponding exponential terms can be estimated and subtracted from the transient, thus permitting identification of lesser terms or corrections to the original estimates. The corresponding transfer function can be written as the sum of terms and then combined into a product.

If the transfer function is known, the transient response to a step input can be written by inspection and the curve plotted. If the transient curve is given, however, the determination of the transfer function is basically a trial-and-error process. Fortunately, most systems have one or two modes dominant so that they can be directly determined with reasonable accuracy. For convenience, the curve obtained in Fig. 11.10 will be used as an illustration in comparison of the direct and inverse processes. Admittedly the answer is already known in this case, but the methods are believed to be applicable to any linear system.

The oscillation is the dominant mode of the transient, but it is not centered about the zero line. Ignoring the zero line, the envelope of the oscillation can be sketched as the dashed lines shown in Fig. 11.11. A dashed median line for the oscillation can then be sketched midway between the two envelope curves.

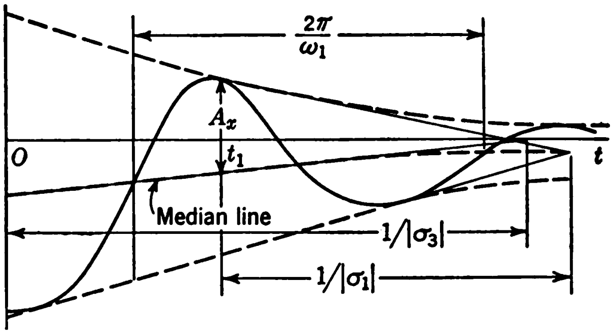


FIG. 11.11. Construction to approximate a given transient response by exponential terms.

The period of the oscillation is measured along the median line, as shown. The time constant of decay of the oscillation is determined by drawing a line from a point on the median line tangent to the envelope, as noted by $1/|\sigma_1|$. The amplitude of the oscillation can be calculated from the single ordinate A_x. The value A_1 at $t = 0$ is then computed from A_x at $t = t_1$.

$$A_x = A_1 e^{\sigma_1 t_1} \cos(\omega_1 t_1 - \phi_1) = A_1 e^{\sigma_1 t_1} \tag{11.26}$$

The ordinate A_x is measured at the peak of an oscillation so that the cosine term is unity.

The median curve itself is another term in the transient response which can be approximated by a single exponential decay. Its time constant $1/|\sigma_3|$ can be determined by sketching a tangent to the median curve and noting the intersection point with the zero axis.

The sum of these two terms from the oscillatory and the median curves can be compared with the actual transient, as shown in Fig. 11.12. The approximation is reasonably exact along the portion of the curve at which the measurements are made; the deviation at the start of the transient can be approximated by the addition of another exponential term with a short time constant $1/|\sigma_4|$.

The complete approximation for the transient in Fig. 11.11 is then:

$$\theta_o = 1 - |A_1|e^{\sigma_1 t}\cos(\omega_1 t - \phi_1) - A_3e^{\sigma_3 t} - A_4e^{\sigma_4 t} \qquad (11.27)$$

The methods of Art. 4.7 can be used to determine the transfer function which would have the above output for a unit impulse input. Or the Laplace transform described in Appendix A could be used to obtain the transform of θ_o denoted as $\Theta_o(s)$

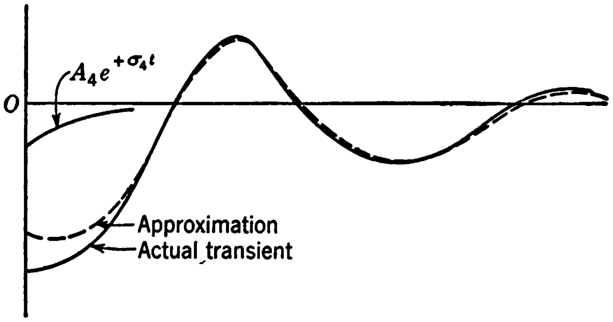


FIG. 11.12. Comparison of actual transient with first approximation.

$$\Theta_o(s) = \frac{1}{s} - \frac{|A_1|}{2}\left[\frac{e^{-j\phi_1}}{s - \sigma_1 - j\omega_1} + \frac{e^{+j\phi_1}}{s - \sigma_1 + j\omega_1}\right] - \frac{A_3}{s - \sigma_3} - \frac{A_4}{s - \sigma_4} \qquad (11.28)$$

The result is the same by either method. The expression in (11.28) could also be expressed as a product of factors, one of which is $1/s$.

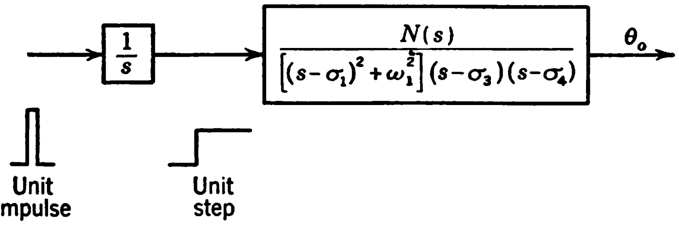

$\frac{1}{s}$
$\frac{N(s)}{[(s-\sigma_1)^2+\omega_1^2](s-\sigma_3)(s-\sigma_4)}$
θ_o
Unit impulse
Unit step

FIG. 11.13. Block diagram of system to produce the transient approximated in Fig. 11.12.

Since the transient in Fig. 11.11 is actually for a unit step input the factor $1/s$ is then dropped to obtain the transfer function of the system (Fig. 11.13). The numerical values of the constants such as ω_1 obtained by fitting the curve of Fig. 11.11 in this article would of course be only approximations to the values in (11.25) from which the curve in Fig. 11.10 was plotted.

11.7. Graphical Solution of Differential Equation

> *The effect of a simple transfer function, such as* $1/(1 + Ts)$, *upon an arbitrary function of time can be approximated quickly by a graphical solution of the differential equation.*

If the transfer function between θ_i and θ_o is $1/(1 + Ts)$, the corresponding differential equation is

$$T \frac{d\theta_o}{dt} = \theta_i - \theta_o \tag{11.29}$$

At any instant, the slope of the output curve projected for a time constant T must equal the difference between the input and the output. This condition can be satisfied graphically (Fig. 11.14).

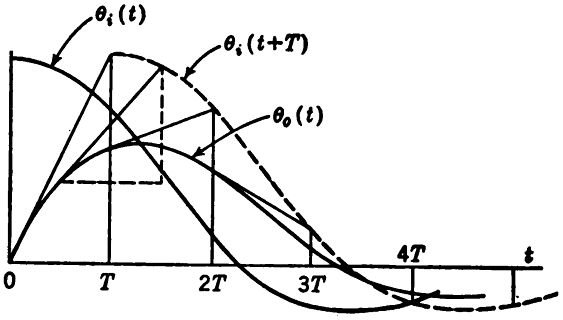


FIG. 11.14. Sketch of the effect of the transfer function $1/(1 + Ts)$ on an arbitrary input.

The time scale is divided into time-constant intervals for convenience, and the input curve is shifted ahead by one interval. The output curve starts at zero and heads toward the initial point of the input curve. The process is continued, with the output curve always being sketched toward the point on the input curve which is one time constant ahead. The differential equation is thus always satisfied.

An approximation to this construction can be obtained by pulling a weight with a string; the length of the string corresponds to the time-constant interval, and the weight traces the output curve if the other end of the string is moved along the input curve. Actually, the horizontal distance between the input curve and the weight should be held constant, although the error does not exceed 10 per cent if the angle of the string does not exceed 25°.

The step-by-step solution of the response of a system can be carried through for several time delays in series just by letting the output of one delay be the input for the next delay. The method is not well adapted, however, to a system having an oscillatory response.

The effect of a single lead can be found graphically rather simply, since the output is determined directly from the input curve at any point.

Digitized by Google Original from UNIVERSITY OF MICHIGAN

Thus, consider the problem of finding the output for a system having the transfer function $1 + Ts$ and the input shown in Fig. 11.15. Any one point on the input curve is extrapolated ahead by the time constant T at the slope of the input curve at that point. For convenience, the curve is divided into sections separated by the time-constant interval. The points thus extrapolated ahead form the pattern for the continuous dashed curve. Each point of the output thus obtained occurs at the same instant of time as the corresponding input point. Thus, the actual output curve is shifted back to the left by the time constant T. Note that the peak of the pulse of the new curve is ahead of the input and that the amplitude is higher than the input.

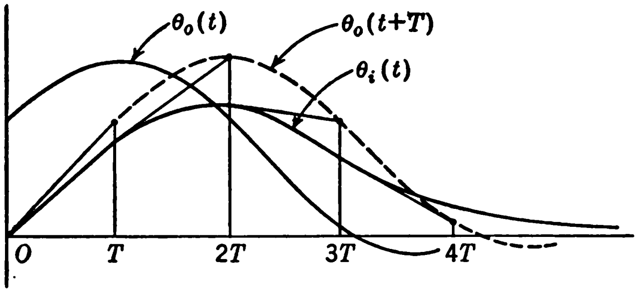


FIG. 11.15. Sketch of the effect of the transfer function $1 + Ts$ on an arbitrary input.

For a system having a lead and lag in combination, it is convenient to divide through to obtain a constant plus a remainder.

$$\frac{1 + T_1 s}{1 + T_2 s} = \frac{T_1}{T_2} + \frac{1 - T_1/T_2}{1 + T_2 s} \tag{11.30}$$

This equation means that the fraction T_1/T_2 of the input would pass without any modification, and the remaining fraction $1 - T_1/T_2$ passes through the time delay $1 + T_2 s$.

The construction for a system having a transfer function with complex poles is considerably more difficult. Any form of transfer function can be handled using convolution, which is discussed in the next article.

11.8. Convolution

Convolution is a process for finding the output of a linear system for any arbitrary input and derives its name from the folding operation involved in the graphical procedure. The input is considered to be a series of impulses, each of which starts the characteristic transient, of the system. This transient, plotted back from the present time, is termed the weighting function, because it weights each impulse of the input as a function of age of the transient in determining the output at any one time.

Many methods can be devised for finding the output of a linear system for an arbitrary input. All are based on the principle of superposition, which is applicable to a linear system because the characteristics of the system are independent of the amplitude of the signals. In finding the output of a particular system for a particular input, attention should be given to exploiting all possible simplifications. The standard procedure, for example, is to approximate the input curve as a series of steps, sketch the outputs for each step, and add the outputs to find the total output. Usually, however, the input curve can be matched more easily with a series of straight-line ramps, and the number of transients introduced is smaller. Another simplification is to plot only the deviation of the output from the assumed input, so that the function being plotted does not build up with time. If a component of the transient is a sine wave, these successive sine waves can be added vectorially. A more general method is to treat the input as a series of impulses; this leads to the concept of a weighting function, which is very useful.

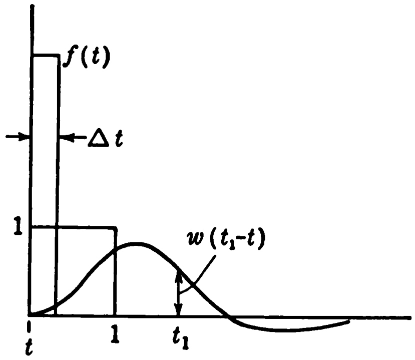


FIG. 11.16. Transient started by a single impulse.

A system is completely specified by its transient response $w(t_1 - t)$ to a unit impulse, as shown in Fig. 11.16. An impulse is the product of a signal $f(t)$ by the time interval dt over which it acts. Such an impulse is a unit impulse if the area involved is unity, or equal to that of the square shown in Fig. 11.16. Such an impulse cannot cause an immediate output but puts energy into the system which gradually decays. An interesting fact is that the area under the output curve is also unity if the d-c value of the transfer function is unity. This fact can be shown from a study of the differential equation

$$\theta_i + a_1 \frac{d\theta_i}{dt} + a_2 \frac{d^2\theta_i}{dt^2} + \cdots = \theta_o + b_1 \frac{d\theta_o}{dt} + b_2 \frac{d^2\theta_o}{dt^2} + \cdots \tag{11.31}$$

Integrating each side of the equation gives the impulse associated with each term.

$$\int \theta_i \, dt + a_1 \int \frac{d\theta_i}{dt} \, dt + \cdots = \int \theta_o \, dt + b_1 \int \frac{d\theta_o}{dt} \, dt + \cdots \tag{11.32}$$

The impulse of the input is $\int \theta_i \, dt$; all the other terms on the left side are zero, because the net changes in θ_i and all its derivatives are zero. The final value of θ_o is also zero; therefore, only the impulse of the output $\int \theta_o \, dt$ remains on the right side of the equation. The net area under the curve $w(t_1 - t)$ in Fig. 11.16 is thus unity.

An arbitrary input curve can be considered to be a series of impulses, each of which starts the characteristic transient of the system. Thus an impulse $f(t)\,dt$ acting at time t causes an output $f(t)\,dt\,w(t_1 - t)$ at time t_1. The total output is the sum of the outputs due to all impulses acting between $t = 0$ and $t = t_1$:

$$\theta_o = \sum f(t)\,dt\,w(t_1 - t) = \int_0^{t_1} f(t)w(t_1 - t)\,dt \tag{11.33}$$

The direct evaluation of the output according to the concept of adding transients would be very tedious, because each impulse must be short compared to the time of response of the system and thus many transients are involved. A slight change in concept, however, produces a very useful method. Note in (11.33) that each ordinate of the input is multiplied by the factor $w(t_1 - t)$, in which $t_1 - t$ is the time interval between the impulse t and the present time t_1. This weighting factor $w(t_1 - t)$ for each impulse can be quickly identified by plotting the transient-response curve back from the present time, as shown in Fig. 11.17.

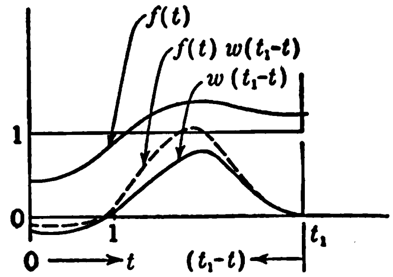


FIG. 11.17. Use of a weighting function to determine an output for any input.

Each ordinate of the input curve can be multiplied by the corresponding ordinate of the weighting function to obtain the dashed curve. The area under the dashed curve is then the value of the output at the time t_1. The numerical value of the area is computed on the basis of the area originally used in determining the transient for a unit impulse. The ordinate of the output times the unit base should thus equal the area under the dashed curve.

To find the output at a later time, the weighting-function curve is plotted back from this time, a new dashed curve is sketched, and the area estimated.

Note that the same output would be produced if the input curve and weighting-function curve were interchanged. This fact corresponds to the concept of interchangeability of blocks, discussed in Art. 4.5. Thus, the input curve can be considered to be the output of a block having a unit-impulse input. Interchanging this block with the block for the system would interchange the nature of the two curves. This fact will be useful in solving the inverse problem of finding the weighting function of a system when given the input and output curves.

11.9. Numerical Convolution

A function of time can be represented by a series of impulses with areas specified by a series of numbers. With the input function and *weighting function thus represented, the process of convolution can be carried out numerically by a process similar to multiplication. The inverse of convolution can then be carried out by a process similar to division.*

The process of convolution involves treating the input function as a series of impulses, each of which starts the characteristic transient of the system. The graphical procedure of Art. 11.8 illustrates that the net output is not affected by interchange of the input curve and transient curve and thus suggests that the transient curve be represented by a series of impulses also. The basic element of the output is due to a single impulse of the input and a single impulse of the transient. For convenience, consider the special case when these impulses are rectangles, each of unit height and unit width. The output can then be shown to be a triangle of unit area (Fig. 11.18).

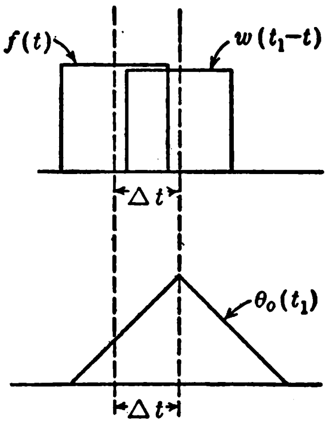


FIG. 11.18. Convolution of two impulses.

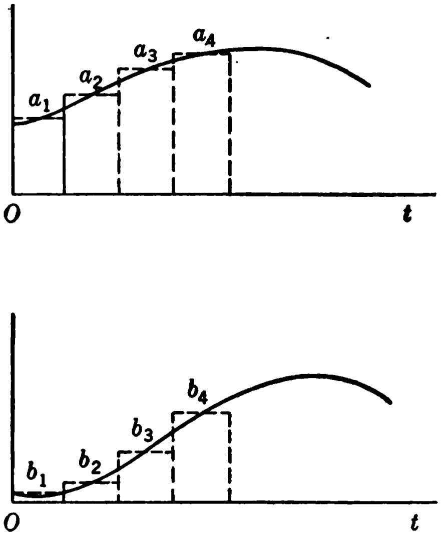


FIG. 11.19. Representation of curves by a series of impulses.

The output when the two rectangles overlap is maximum and is unity. As the rectangles are displaced, the output falls off linearly, reaching zero for a unity displacement in either direction. The area of the triangle is half the base times the altitude, or unity.

The impulse concept suggests a criterion for approximating a general input curve a or weighting-function curve b as shown in Fig. 11.19. The curves are divided into a series of equal time intervals; a series of rectangles are then sketched, with ordinates $a_1, a_2, a_3, \ldots$ and $b_1, b_2, b_3, \ldots$ such that the area of the rectangle for each interval is equal to the area under the actual curve. Considering the time intervals to be unity, the ordinates are then numerically equal to the impulses.

The impulses in the output can be located at the proper time interval by the process shown in Fig. 11.20. The row of numbers a_1b_1, a_1b_2, a_1b_3, . . . represents the impulses in the output due to the first impulse a_1 of the input and all the impulses b of the transient. The next row is the impulses in the output due to the second impulse a_2 of the input; the offset of rows makes all the impulses in one column occur in the same time interval; thus

$$\begin{array}{ccccccc}
b_1 & b_2 & b_3 & \bullet & \bullet & \bullet & \\
a_1 & a_2 & a_3 & \bullet & \bullet & \bullet & \\
\cline{1-3}
a_1b_1 & a_1b_2 & a_1b_3 & \bullet & \bullet & \bullet & \\
 & a_2b_1 & a_2b_2 & a_2b_3 & \bullet & \bullet & \bullet \\
 & & a_3b_1 & a_3b_2 & a_3b_3 & \bullet & \bullet \\
\cline{1-3}
c_1 & c_2 & c_3 & & \bullet & \bullet & \bullet
\end{array}$$

FIG. 11.20. Numerical convolution.

$$c_3 = a_1b_3 + a_2b_2 + a_3b_1 \tag{11.34}$$

where the c's are the impulses of the output. All parts of c_4 and beyond are not shown. The output is constructed as a series of rectangles with ordinates c_1, c_2, c_3, . . . , and then a smooth curve is sketched through the tops of the rectangles, making the area of a rectangle and that under the corresponding part of the curve match.

The inverse of convolution involves finding the input curve to produce a given output for a system with a known weighting function. The process can be illustrated by using the same symbols as before (Fig. 11.21).

$$\begin{array}{r|cccccc}
 & a_1 & a_2 & \bullet & \bullet & \bullet & \\
\cline{2-4}
b_1\ b_2\ b_3 \bullet \bullet \bullet & c_1 & c_2 & c_3 & \bullet & \bullet & \bullet \\
 & a_1b_1 & a_1b_2 & a_1b_3 & \bullet & \bullet & \bullet \\
\cline{2-4}
 & 0 & a_2b_1 & c_3-a_1b_3 & \bullet & \bullet & \bullet \\
 & & a_2b_1 & a_2b_2 & \bullet & \bullet & \bullet \\
\cline{3-4}
 & & 0 & a_3b_1 & \bullet & \bullet & \bullet
\end{array}$$

FIG. 11.21. Inverse of convolution.

The first impulse c_1 of the output is divided by the first impulse b_1 of the weighting function to determine the first impulse a_1 of the input. The entire transient produced by this first impulse of the input, a_1b_1, a_1b_2, a_1b_3, . . . , is then subtracted from the corresponding impulses of the output to give the remainder of the output yet to be accounted for. In the next row the symbol a_2b_1 is given for the difference $c_2 - a_1b_2$ because this impulse can be due only to the second impulse a_2 of the

Digitized by Google Original from UNIVERSITY OF MICHIGAN

input and the first impulse b_1 of the weighting function. The process is continued until the entire input function is determined.

The explanation above has assumed that the weighting function is known; the same process can be used, however, to determine the weighting function if the input and output functions of time are known. Physically, subtracting the first row in Fig. 11.21 would correspond to subtracting all the impulses in the output caused by all the impulses in the input and the first impulse of the weighting function.

The convolution process indicates the concepts involved in working directly with functions of time. The solution of any given problem, however, is apt to be more tedious unless some form of mechanization of the routine is employed. These methods constitute a separate study in themselves and so are not treated in this book.

CHAPTER 12

GRAPHICAL ANALYSIS OF NONLINEAR SYSTEMS

The purpose of a graphical analysis of nonlinear systems is to develop some understanding of such common effects as coulomb friction, pick-off dead zone, and backlash of gears. These nonlinearities are simple in that they are all on-off effects; no arbitrary curve, such as the saturation curve of a magnetic field, is involved. The system is therefore linear within a definite range of operation; the only new problem is to determine quickly the points of discontinuity.

A quadratic system is studied because only two variables are involved: displacement and velocity. A so-called "phase-plane" diagram of velocity versus displacement permits switching points to be determined readily and the initial conditions for each new transient to be automatically satisfied.

The frequency-response method can be extended to apply to nonlinear components if the distorted output is approximated as the fundamental of a Fourier series. The higher harmonics are neglected, on the assumption that they are attenuated in other elements around the loop.

A comparison of results of the phase-plane diagrams and frequency-response plots for simple systems shows good agreement regarding all effects studied.

12.1. Sketch of Time Response

> *A system which is linear except for an on-off error detector has a response which is a series of linear transients. The problem is to determine the switching points.*

A general single-loop system with an on-off pick-off is shown in Fig. 12.1. Consider the problem of sketching the transient response of this system using simplifications, such as treating the load as a pure inertia and neglecting the spring. The motor has a time constant T in coming up to speed for a suddenly impressed voltage; the relay contact circuit causes the motor speed to decay with the same time constant when neither relay is energized.

If the input is suddenly moved through a large angle, closing the circuit of the forward, or F, relay, the motor speed and angle curves are as shown in Fig. 12.2. The speed curve v_o approaches a maximum velocity of V_m exponentially; the angle curve θ_o approaches a constant slope which is offset along the time axis by the time constant T.

At the dead zone in Fig. 12.3 the forward relay drops out, and a new transient is started. The speed starts to decay toward zero, and the angle continues to increase, with its maximum possible travel being the motor speed V_A times the time constant T. Actually, at point B, the reverse, or R, relay picks up and starts another transient. The process could be continued, but this example demonstrates that the output angle θ_o must be known in order to determine a switching point; the velocity must be known to match the end of one interval with the start of the next.

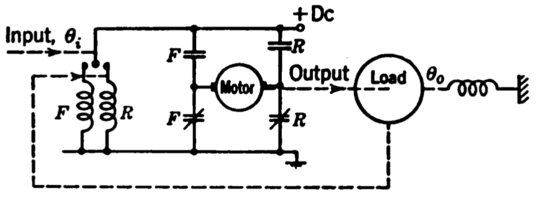


FIG. 12.1. A general single-loop on-off system.

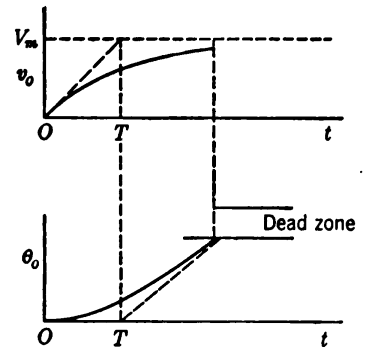


FIG. 12.2. Sketch of initial portions of transient response.

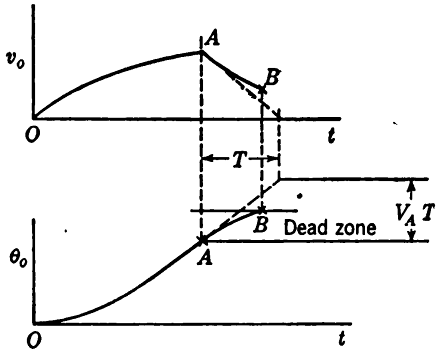
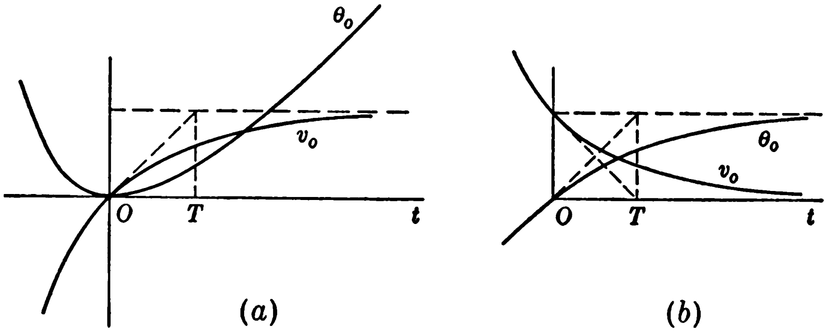


FIG. 12.3. Transient within the dead zone of the pick-off.



(a) (b)

FIG. 12.4. Standard curves of velocity and angle versus time: (a) with motor excited; (b) with motor not excited.

The construction could be made more rapidly by use of a standard set of curves which could be used as templates (Fig. 12.4). The first set of curves is for the condition when voltage is applied to the motor so that the steady-state condition is constant velocity; the second set applies to the decay interval in the dead zone of the contacts. Either set of curves could be entered at any point and the next section of the transient traced. The plot of any one section of the transient can thus be made quickly, but the number of switching points involved makes the use of a plot of velocity versus displacement preferable, as shown in the next section.

12.2. Phase-plane Diagrams

A plot of velocity versus displacement of the output describes the system for each region of operation. Switching points are immediately apparent, and initial conditions are automatically satisfied.

If the standard curves shown in Fig. 12.4 were replotted as velocity versus angle, they would appear as shown in Fig. 12.5. The direct way

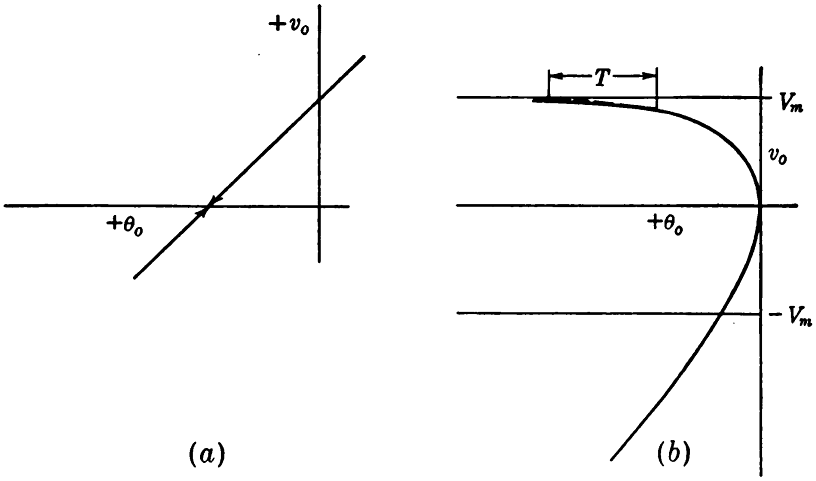


FIG. 12.5. Curves of velocity versus angle: (a) with motor not excited; (b) with motor excited.

to obtain these curves is to select a value of time in Fig. 12.4, read the corresponding values of both angle and velocity, and plot these values in Fig. 12.5. The limiting conditions on these curves can be sketched directly. The plot of velocity versus angle in the dead zone (Fig. 12.5*a*) is a straight line, because both variables approach their final values in the same manner. Thus, in one time-constant interval each variable decays to $1/e$ of its previous value; *i.e.*, the ratio of velocity to the remaining angle change is constant. The slope of the line can be made unity by properly selecting the ratio of scales on the θ_o and v_o axes.

The curve for the case of positive voltage impressed across the motor

(Fig. 12.5*b*) has a slope of 2 at negative maximum velocity. The curve approaches the maximum positive velocity asymptotically. The θ_o axis can be considered to be a time axis because the speed is nearly constant. The net speed change remaining thus approaches zero exponentially on the plot, as indicated by the tangent in Fig. 12.5*b*.

The curves in Fig. 12.5 describe angle changes as a function of velocity, but the initial value of angle can be offset by any arbitrary amount. This fact permits a family of curves to be drawn with different horizontal offsets, as shown in Fig. 12.6. The curves for negative voltage impressed on the motor are on the left side of the figure; the curves for positive voltage are on the right. Any transient can be sketched by starting at an arbitrary point and following the pattern of the curves. Thus, a sudden displacement of the input can make the initial error at point *A*, which causes the motor to drive to nearly full speed before entering the dead zone, point *B*. The momentum of the load causes a single overshoot, with the transient ending at point *E* with the velocity zero in the dead zone.

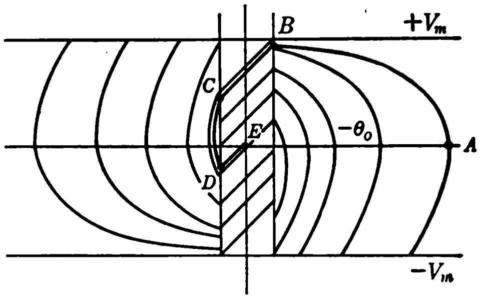


FIG. 12.6. Phase-plane curves for the system of Fig. 12.1 without spring load.

A quadratic system thus permits the characteristics of the system to be described by a set of curves from which any transient can be traced for any arbitrary initial condition. A continuously changing input, however, would require that the position of the dead zone be shifted continuously. The time involved in motion of the output would also have to be known in order to establish the time of entering the dead zone.

12.3. Resonant System with Coulomb Friction

The phase-plane diagram for a resonant system is a circle. Coulomb friction shifts the center of the circle with each change in velocity, causing the amplitude of oscillation to decrease by a constant amount.

The system shown in Fig. 12.1 can be solved very easily for the special case when the damping of the motor is replaced by coulomb friction and the spring action on the load is considered. The first step is to establish the curves of velocity versus displacement of the load with the motor not energized and the coulomb friction neglected. The remaining spring-mass system has a natural undamped oscillation. The displacement and velocity curves are thus sine waves which are displaced 90° in phase. A sine wave can be obtained as the projection of a rotating vector; the 90°

phase difference can be obtained by separating the axis of projection by 90°, as shown in Fig. 12.7.

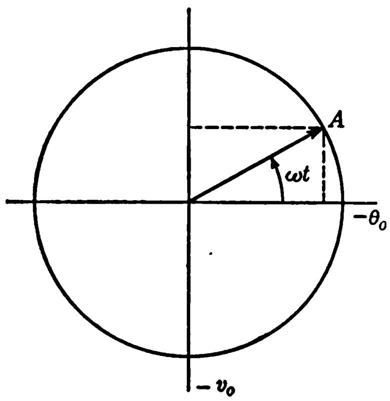


FIG. 12.7. Justification of a circle as the phase-plane curve for a resonant system by projection of a rotating vector on two axes at right angles.

The projection of the vector on the vertical axis is velocity, and the projection on the horizontal axis is displacement. These conditions make the circular path of the head of the vector be the desired phase-plane curve. The angle of the vector is ωt, which permits time to be read from the diagram in terms of angle divided by natural frequency. Any oscillation, once started, would continue at constant amplitude. A set of concentric circles thus describes the behavior of the spring-mass system itself. The effect of energizing the motor is to shift the center of the oscillation by the steady-state deflection of the spring produced by the motor torque. The phase-plane diagram for the system in Fig. 12.1 with relay action included is thus as shown in Fig. 12.8*a*.

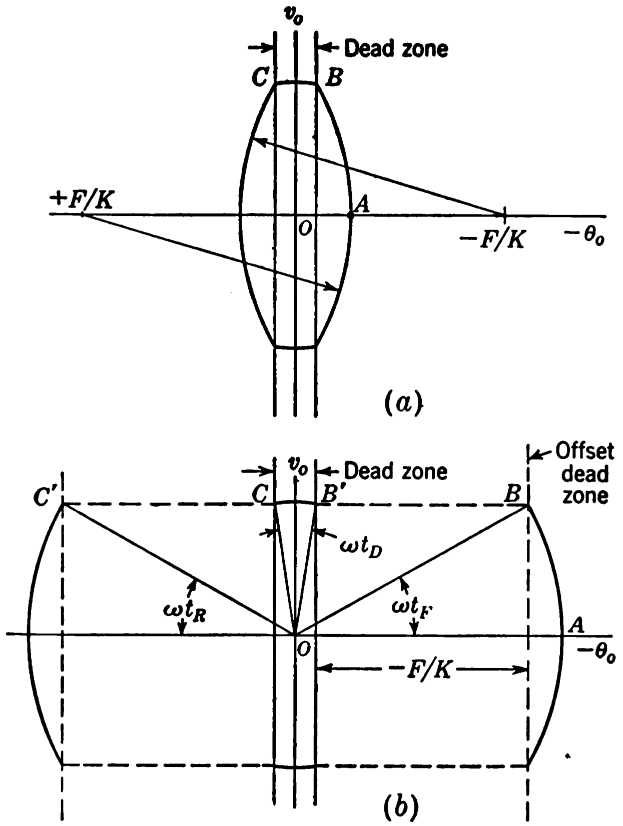


FIG. 12.8. Phase-plane diagrams for system in Fig. 12.1 with spring load and no damping: (*a*) circles offset, owing to motor torque; (*b*) offset dead-zone plot.

The relay dead zone is established with the center of the circles at the center of the dead zone for the condition of the motor being deenergized. At the left of the dead zone, however, the F relay circuit is closed. The motor torque shifts the center of the circles to the right by F/K, where F is the motor torque and K is the spring constant. Similarly, operation to the right of the dead zone involves the center of the circles being shifted to the left by F/K.

Another form of presentation involves keeping the center of the circles fixed but shifting the position of the dead zone, as shown in Fig. 12.8*b*. The circle arcs in this figure are exactly the same as those in Fig. 12.8*a*. The dashed lines from B to B' and from C to C' represent the instantaneous shifts in the apparent position of the dead zone with respect to a fixed center of oscillation. This construction is perhaps less natural than the former one, but it does permit a more rapid estimate of the time involved per cycle. Thus, the angles representing ωt all have a common center; the subscript F is for forward, D for dead zone, and R for reverse. The period of oscillation for the servo system compared to the load itself is reduced by the fraction $\omega(t_F + t_D + t_R)/\pi$. Coulomb friction is assumed to produce a constant force on the load opposing the existing motion. The response of the load without servo action is then as shown in Fig. 12.9.

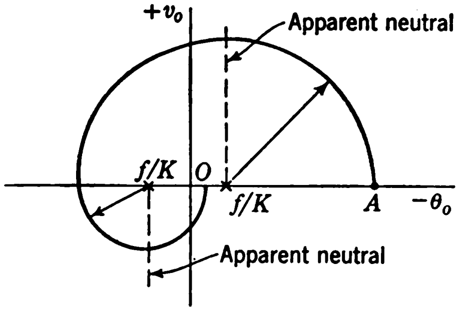


FIG. 12.9. Effect of coulomb friction.

Start with the load displaced at the point A; the friction force causes the apparent neutral position of the spring to be shifted to the right by f/K. After a semicircle, the velocity reverses sense, and the center of oscillation now shifts to the left of the true neutral of the spring by f/K. The oscillation stops after this semicircle because the spring force is less than the coulomb friction force.

A system with both relay action and coulomb friction would involve the same construction as before with the exception that the shift of the center of oscillation would occur both at the dead zone and at a change in velocity.

12.4. Resonant System with Viscous Damping

A damped oscillation can be represented by a rotating vector of decreasing length, so that the tip of the vector traces a logarithmic spiral. This spiral is the phase-plane curve if the velocity axis is tilted with respect to the vertical by the proper angle.

Digitized by Google Original from UNIVERSITY OF MICHIGAN

The differential equation for the load with inertia J, viscous damping D, and spring constant K is,

$$Kx + D\frac{dx}{dt} + J\frac{d^2x}{dt^2} = 0 \tag{12.1}$$

Assuming a solution of the form Ae^{st} leads to an algebraic equation in s for which the roots are

$$s = -\frac{D}{2J} \pm \sqrt{\left(\frac{D}{2J}\right)^2 - \frac{K}{J}} = \sigma \pm j\omega = \omega_0\underline{/\pm\psi^\circ} \tag{12.2}$$

The transient response is thus a damped sine wave, which can be represented by the projection of a rotating vector whose length is continuously decreasing (Fig. 12.10). The tip of the vector traces a logarithmic spiral whose radius decreases by the constant factor $\epsilon^{\sigma(2\pi/\omega)} = 0.14$ for each revolution, corresponding to a root whose damping ratio $\zeta = 0.3$ and whose angle ψ° is $\cos^{-1}\zeta$. The instantaneous value of displacement is the projection of the vector on the positive real axis. Velocity changes are ahead of displacement changes by the angle ψ°, as shown. This relationship can be checked by noting that the vector at point A represents the maximum displacement, because the projection line is tangent to the curve and the velocity projection is zero. The angle at the center O is a measure of the elapsed time for this spiral just as for the circle.

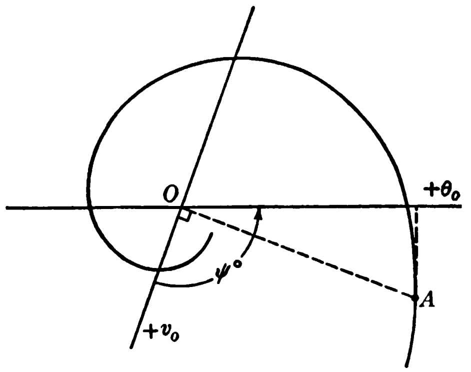


FIG. 12.10. Logarithmic spiral for a damped resonant system with projection axes for angle and velocity separated by an oblique angle.

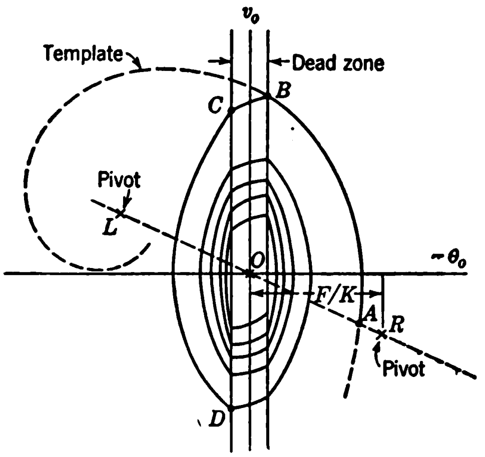


FIG. 12.11. Phase-plane diagram with shift of point of convergence of spiral due to motor torque.

Consider now the construction to find the response of the system as a servo. When a relay contact is picked up and full voltage impressed across the motor, the equilibrium position of the load is shifted by the amount of the spring displacement under full motor torque. The shift is along the inclined axis, because the velocity cannot instantaneously change at the time of switching (Fig. 12.11). Starting at the point A, the transient is a logarithmic spiral about the point L. A template, indicated by the dashed curve, can be used for tracing this curve. When the dead zone is reached at point B, the new equilibrium position is shifted to point O. The template can be shifted to this new position with center O as a pivot to rotate the template until the curve touches point B. After tracing the curve to point C, the template is shifted to point R as a pivot and the curve traced from C to D. This process can be continued until the load reaches zero velocity within the dead zone.

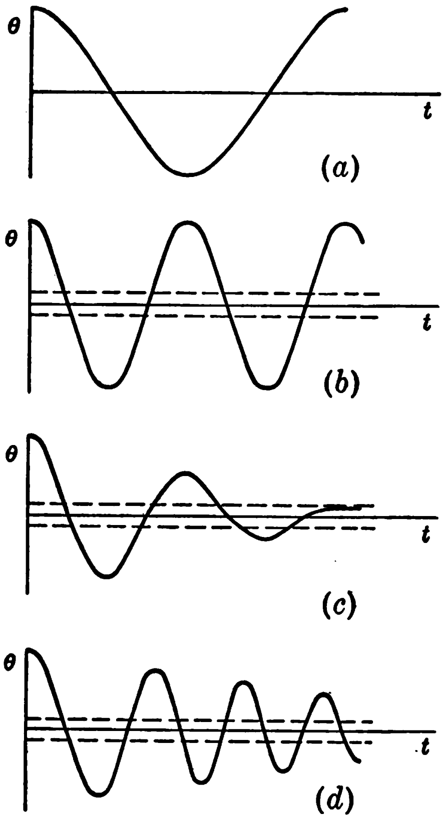


FIG. 12.12. Comparison of transient responses: (*a*) undamped resonant system without servo; (*b*) undamped system with servo; (*c*) system with coulomb friction; (*d*) system with viscous damping.

A comparison of results is shown in Fig. 12.12. The undamped oscillation of the resonant load serves as a basis for comparison with other results. Servo operation (Fig. 12.12*b*) decreases the period of an oscillation without decreasing the amplitude. The addition of coulomb friction or viscous damping causes the oscillation to die out.

12.5. Relay Time Delay

Relay time delay causes an oscillation to be maintained at a small amplitude. A counteracting effect can be achieved by applying reverse voltage on the motor as soon as the forward relay drops out.

A typical small relay will have a time delay in picking up or dropping out of the order of 5 msec. For a system whose natural frequency is 5 cps, this time delay is about one-sixth of the time for 1 rad of oscillation and might well be expected to have an appreciable effect on the response. The phase-plane construction for the system in Fig. 12.1 is basically the same as that in Fig. 12.11, except that relay time delay shifts the point of switching (Fig. 12.13*a*).

The transient starts at the right and continues to the dead zone, at which time the relay is deenergized. The time delay before dropout causes the motor to remain energized for a time corresponding to a 10° angle on the logarithmic spiral. The pivot point of the template is then shifted to the center of the dead zone. The template is then traced beyond dead zone by 10°, to allow for relay pickup time, before the pivot point is shifted again. The construction is repeated for the remaining cycles shown.

Note that the decrease in amplitude becomes less on each cycle; thus the system might settle down to continuous oscillations at an amplitude which is slightly smaller than that of the last cycle shown. An interesting way to check this possibility is to start a transient at an amplitude which is less than the suspected equilibrium amplitude and note whether the oscillation will build up (Fig. 12.13*b*). The amplitude does build up with successively smaller increases. The limiting cycle of oscillation must be between the two limits, build-up in Fig. 12.13*b* and convergence in Fig. 12.13*a*.

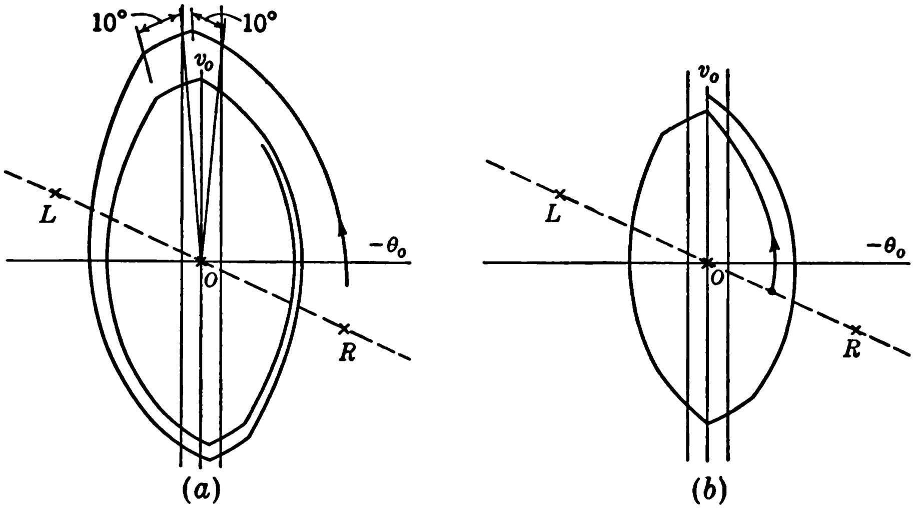


FIG. 12.13. Effect of relay time delay: (*a*) convergence of large-amplitude oscillation; (*b*) build-up of small-amplitude oscillation.

Relay time delay is thus the first of the effects studied which causes an oscillation to maintain itself in spite of damping on the load. The delay in relay dropout causes the motor to continue to accelerate the load after the dead zone is entered. The delay in relay pickup allows the load to overshoot the dead zone before motor plugging action takes place. For very large amplitude oscillations, these effects are more than counterbalanced by the damping forces on the load. At small amplitudes, however, the increase in velocity of the load due to relay-drop-out delay becomes greater, and the velocity lost per cycle due to damping becomes less. The oscillation reaches an equilibrium at an amplitude such that the velocity increase per cycle is equal to the velocity decrease per cycle.

The study of the effect of relay time delay suggests a change which should have a counteracting effect. Plugging action of the motor could be started as soon as one relay drops out rather than being delayed until the other relay picks up (Fig. 12.14). The relays F and R control the operation of the actual motor relays F' and R'. The control is such that R' is picked up as soon as F drops out, similarly F' is picked up as soon as R drops out. This advance pickup is effective for only a short time interval during which the contact of R, for example, would in turn cause R' to hold in. If R is not picked up, then R' drops out after this short interval, allowing the system to come to rest.

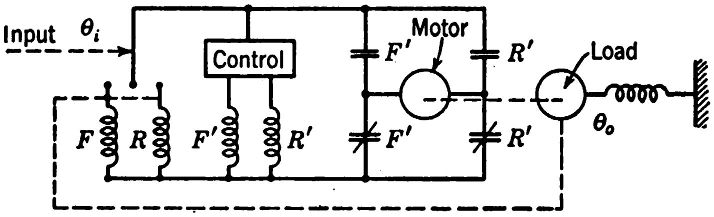


FIG. 12.14. Circuit to achieve advance motor plugging.

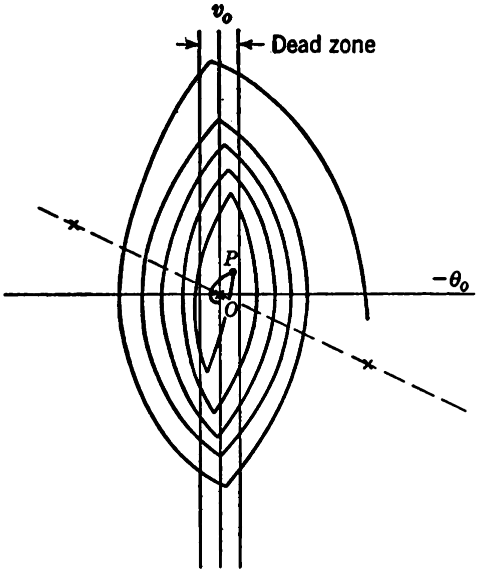


FIG. 12.15. Phase-plane diagram for system with advance plugging.

The corresponding transient is shown in Fig. 12.15. When the dead zone is entered, the reverse relay is energized and picks up after its own time delay. The plot is shown for the case in which the relay still picks up before contact on the far side of the dead zone is made.

12.6. On-Off Effects in Linear Quadratic Systems

Many nonlinear characteristics can be approximated by straight lines, making the system linear within several ranges of amplitudes of signals. The phase-plane curves for each range of operation are often already known, so that the complete construction can be made very rapidly.

The actual characteristic of a pick-off and amplifier combination is shown as part of the block diagram of the system in Fig. 12.16. The straight-line approximations to this characteristic define three ranges of operation of the system, each of which is linear. The corresponding phase-plane diagram (Fig. 12.17) can be developed in a series of simple steps.

First the three regions of operation are established by vertical lines. Each of these ranges represents a system which is discussed in a previous article. The dead-zone region $|\theta_e| < \theta_d$ is described by straight lines on the phase plane. Similarly, the saturation region is the same as that for a relay being energized. Finally, the linear range has motor torque proportional to angle change.

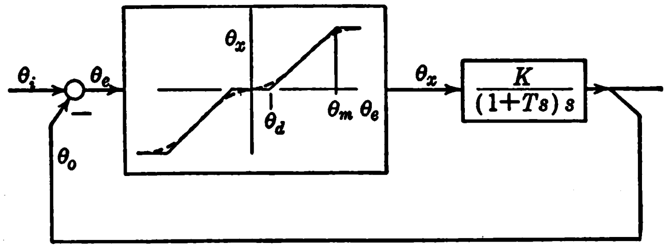


Fig. 12.16. Block diagram with curve to denote flat spot and saturation of the pick-off and amplifier.

An interesting aspect of this diagram is that the slopes of the curves do not change at the boundaries of the various ranges. Slope is associated with acceleration because the θ_o axis can be considered to be a time axis, once a velocity is given. Acceleration of the load due to the voltage applied to the motor has no discontinuities; therefore, the slopes of the curves will have no discontinuities.

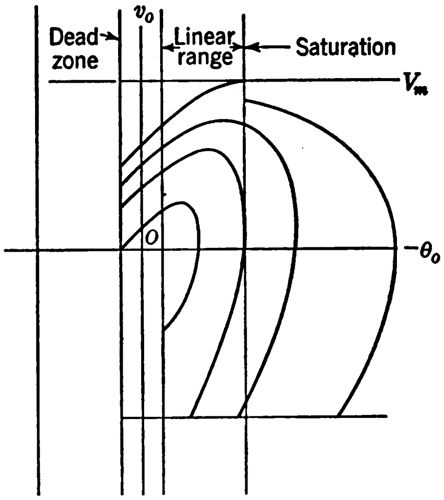


Fig. 12.17. Phase-plane diagram showing the various regions of operation due to characteristic shown in Fig. 12.16.

The phase-plane diagram is limited to a quadratic system; thus, the number of forms of velocity-displacement curves is limited. Three curves are shown in Fig. 12.17, and three more arise if the transfer function $K/(1 + Ts)s$ is replaced by K/s^2. Other on-off effects can be added to the given system by further dividing the operating ranges. Coulomb friction, for example, would make the curves for positive velocity differ from those for negative velocity. Backlash is a special on-off effect and is treated in the next topic.

12.7. Backlash

Backlash is a destabilizing effect, because it permits the motor to acquire a velocity at the reversal point before the output is changed.

Lowering the gain of the loop decreases the amplitude of oscillation, and coulomb friction may stop the oscillation.

Backlash is somewhat more difficult to analyze than the other on-off effects, because two angles are involved. The motor angle θ_M takes up the backlash in the gears before the load angle θ_o must follow. The inertia of the load J_L is not connected to the motor during this interval, which requires that the block diagram be as shown in Fig. 12.18.

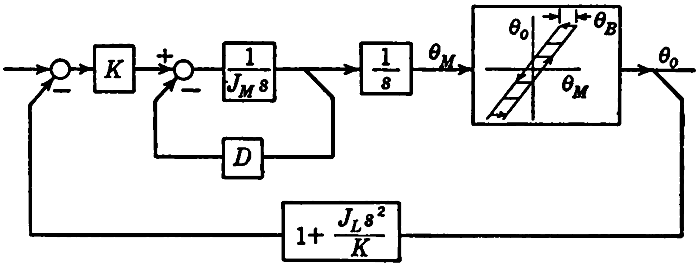


FIG. 12.18. Block diagram of system with backlash.

The operation of the system is illustrated for one reversal of the motor in Fig. 12.19. The motor leads the load by half the backlash up to the point of reversal. The motor then accelerates through the backlash angle without motion of the load. This acceleration is essentially constant, is proportional to the fixed angle of the load, and is limited only by the inertia of the motor; $\alpha = K\theta_o/J_M$. The velocity acquired by the motor in rotating through the backlash angle θ_B is

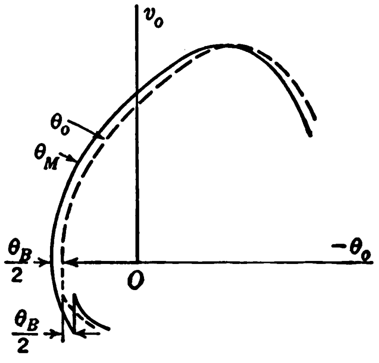


FIG. 12.19. Phase-plane diagram showing the effect of backlash.

$$v_M = \sqrt{2\alpha\theta_B} = \sqrt{\frac{2K\theta_o\theta_B}{J_M}} \qquad (12.3)$$

Assuming pure impact at the time of remeshing of the gears, the total momentum of the motor and load is then equal to that of the motor alone before impact. The output thus suddenly acquires a velocity

$$v_o = \frac{J_M v_M}{J_L + J_M} = \frac{\sqrt{2K\theta_o\theta_B J_M}}{J_M + J_L} \qquad (12.4)$$

A phase plane could be constructed with this velocity plotted as a function of the output angle at the point of reversal. The stability of the system alone, however, can be determined from a simpler diagram in which the velocity gains and losses per cycle are compared (Fig. 12.20). The velocity gain due to backlash is proportional to the square root of the amplitude of oscillation. The decrease in velocity due to damping during linear operation is proportional to the amplitude of oscillation, as shown by the straight line. The intersection at point A

Digitized by Google Original from UNIVERSITY OF MICHIGAN

determines the amplitude of continuous oscillation. Lowering loop gain would cause less voltage to be impressed on the motor for a given amplitude, and thus less velocity of the load would be acquired during the backlash interval. Lowering the gain would also increase the damping ratio of the natural oscillation of the system, so that the velocity lost per cycle would increase. Both of these effects therefore cause the point of intersection in Fig. 12.20*a* to move to a stable operating point of lower amplitude.

For small amplitudes of oscillation the effect of coulomb friction is very important. Friction reduces the amplitude or velocity of an oscillation by a fixed amount each cycle and shifts the straight line up, as shown in Fig. 12.20*b*. If the friction were present on the motor side, it would decrease the torque available for acceleration during the backlash interval and therefore offset the parabola to the right. The combined effect of coulomb friction can therefore make the velocity lost per cycle always greater than the velocity gained, for any amplitude of oscillation.

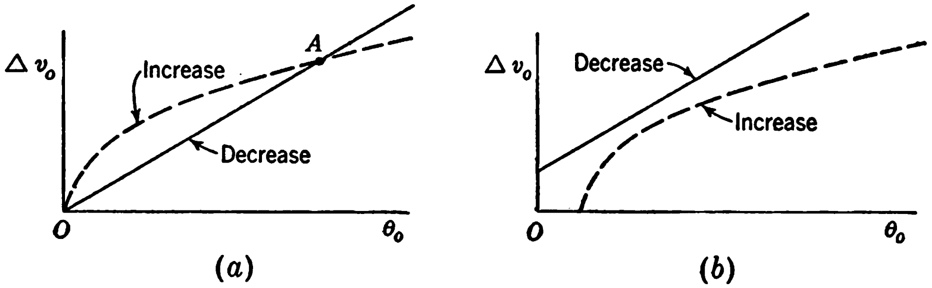


FIG. 12.20. Effect of amplitude of oscillation on the velocity changes per cycle due to backlash and damping: (*a*) equilibrium condition at point *A*; (*b*) effect of coulomb friction.

The concepts developed in the study of quadratic systems could be extended to more complicated systems. Such extensions, however, are always subject to error and therefore must be used with caution. Fortunately, the frequency-response method can be extended to nonlinear systems in approximate form, as shown in the following articles. Finally, a comparison of results of the phase-plane diagrams and the frequency-response plots for simple systems indicates complete agreement qualitatively.

12.8. Frequency Response of Nonlinear Components

> *A nonlinear component produces distortion of its transmitted signal; hence the amplitude of the output signal must be considered. However, components of the output above the fundamental are neglected. The phase shift and attenuation are then functions of amplitude of signal and are often independent of frequency.*

The frequency-response method is based on the description of a system by its transmission of sinusoidal signals of different frequencies. A linear component has a sinusoidal output whose phase shift and attenuation describe the effect of the component on the signal. A nonlinear component, however, has a distorted output and requires that the fundamental component of the output be considered in defining phase shift and attenuation. The higher harmonics are neglected because of the complexity of doing otherwise. The justification is that high-frequency signals are usually attenuated elsewhere in the servo system. The results are only approximate, however; therefore further approximations are justified for convenience in obtaining results rapidly.

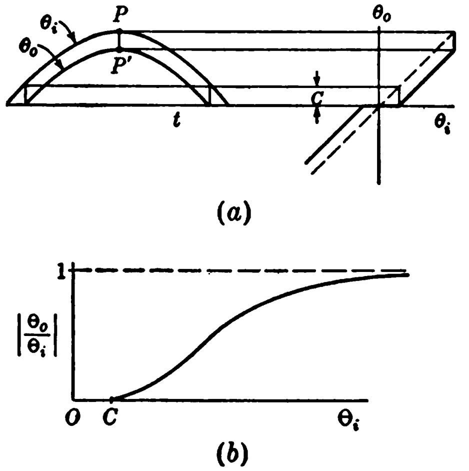


FIG. 12.21. Frequency response versus amplitude for a pick-off having a flat spot at low amplitude: (*a*) construction method; (*b*) attenuation versus amplitude.

Consider first the output of an amplifier which has a dead zone in its characteristic for low-amplitude signals (Fig. 12.21*a*). A point *P* on the input curve is projected horizontally to the 45° line and then down to the amplifier characteristic. This point is projected back on the output to the same time base as the input curve, at point *P'*. Note that each point of the output is less than the input by the amount *C*, except for an input less than *C*, for which the output is zero. The construction is independent of frequency, and no phase shift is introduced because of the symmetry of the diagram.

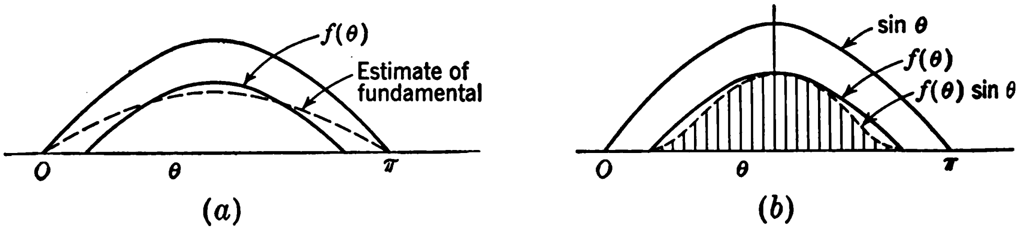


FIG. 12.22. Determination of the fundamental in a distorted output: (*a*) visual estimate; (*b*) use of weighted curve.

Attenuation is a function of the amplitude of the signal compared to the offset *C*, as sketched in Fig. 12.21*b*. The ratio of output to input is zero for θ_i less than *C* and then builds up gradually for larger signals, finally approaching unity for very large signals. A simple way to estimate the fundamental component of the output is merely to sketch a sine wave on top of the output wave (Fig. 12.22*a*). The first tendency is probably to make the area under the sine curve equal to that under the output curve. This result is not correct, however, in that the ordinates of the output should be weighted sinusoidally, as given by the formula

$$A = \frac{1}{\pi} \int_0^\pi f(\theta) \sin \theta \, d\theta \tag{12.5}$$

A direct evaluation of the formula involves sketching a new curve for the weighted value of the output and estimating the area under it (Fig. 12.22*b*).

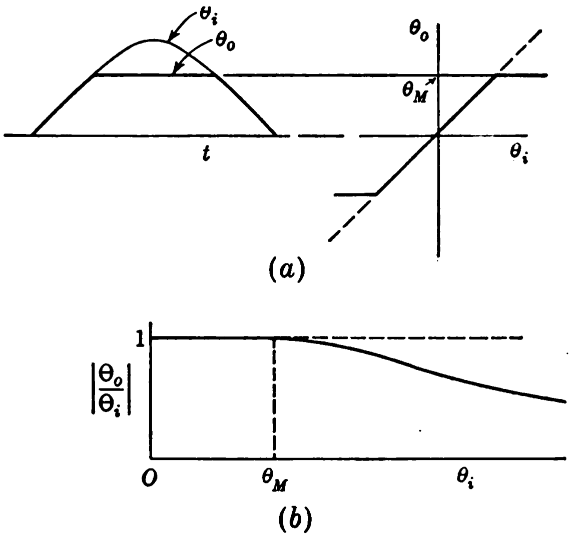


FIG. 12.23. Frequency response of an amplifier having saturation: (*a*) construction method; (*b*) attenuation versus amplitude.

Another common nonlinear characteristic is saturation, as shown in Fig. 12.23*a*. The construction is independent of the frequency of the signal, and no phase shift is involved. Attenuation is a function of amplitude, as indicated in Fig. 12.23*b*. For input signals above the saturation limit, the output remains nearly constant; therefore, the output-to-input ratio falls off hyperbolically.

The pick-off for the relay servo (Fig. 12.1) has the nonlinear characteristic shown in Fig. 12.24*a*. The attenuation characteristic is shown in Fig. 12.24*b*. The presence of time delay in the relay pickup and dropout shifts the output square wave along the horizontal axis by a fixed time. The fundamental of the output is then shifted in angle in proportion to the frequency of the input wave.

The hysteresis type of characteristic associated with backlash has the output angle for a sinusoidal motor angle shown in Fig. 12.25*a*. This

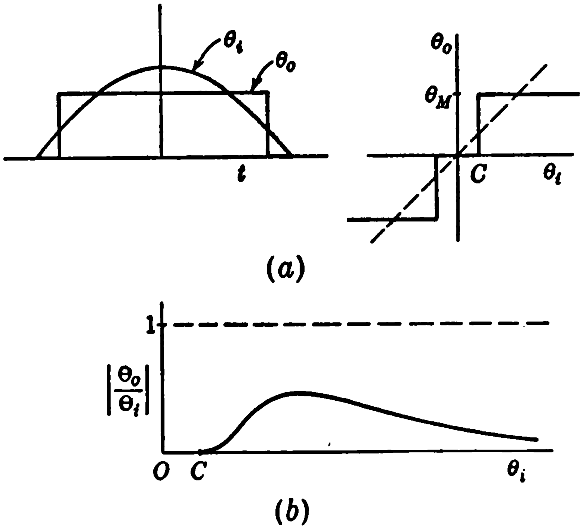


Fig. 12.24. Frequency response of relay-contact-type pick-off: (*a*) construction method; (*b*) attenuation versus amplitude.

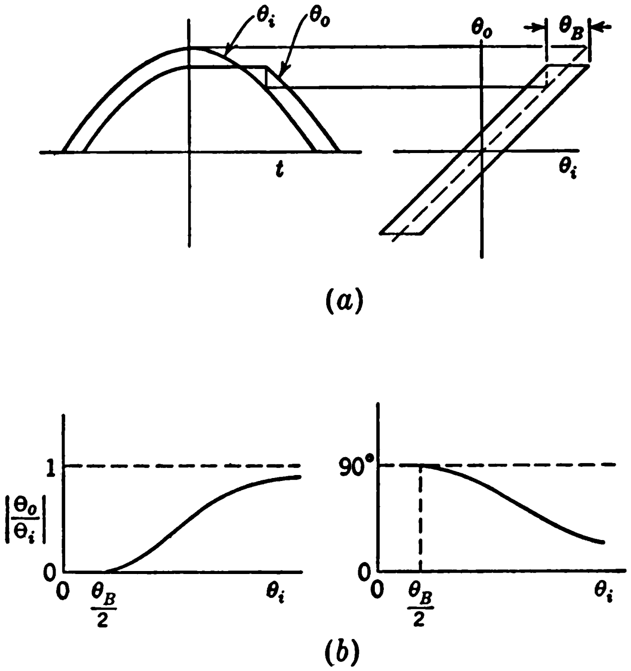


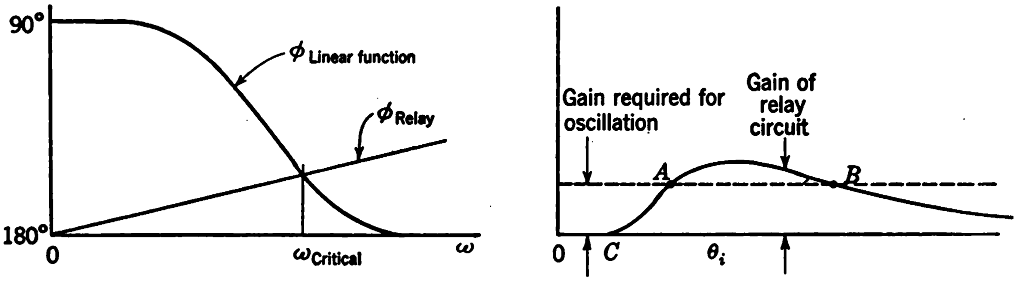
Fig. 12.25. Effect of backlash on frequency response: (*a*) construction method; (*b*) phase shift and attenuation versus amplitude.

diagram is independent of the frequency of the signal. Phase shift and attenuation are both functions of amplitude of signal, however, as shown in Fig. 12.25*b*. For very large signals backlash has no effect, but for smaller signals the phase shift and attenuation both become significant, approaching 90° shift and complete attenuation for an input signal equal to half the backlash angle.

12.9. Frequency Response of Nonlinear Systems

The frequency response of a nonlinear component can be combined with the frequency response of the remainder of the loop in order to determine the limits of stability of the system. The results for quadratic systems agree with the results obtained by phase-plane diagrams.

The analysis of relay time delay is made with a system whose linear portion has a transfer function of the form $K/(1 + Ts)s$. The frequency response of this part of the system can be combined with that of the relay itself, as shown in Fig. 12.26.



(*a*) (*b*)

FIG. 12.26. Determination of equilibrium operation of system with relay time delay by intersection of frequency-response curves: (*a*) frequency set by phase shift; (*b*) amplitude set by attenuation.

The phase shift of each part of the system is a function of frequency only (Fig. 12.26*a*). The phase shift due to the relay is plotted up from 180°, so that the intersection point with the curve for the linear part of the system determines the frequency for which the total phase shift is 180°. The attenuation characteristics are a function of amplitude alone (Fig. 12.26*b*). Setting the gain for the linear part of the system establishes the gain which is required of the relay for the loop gain to be unity. The intersection points *A* and *B*, therefore, represent possible amplitudes of oscillation. Operation at point *B* corresponds to the limit cycle determined in Fig. 12.13. The point *A* does not correspond to an equilibrium condition, because a slightly larger amplitude would continue to build up whereas a slightly smaller amplitude would decay.

The block diagram for the system having backlash is given in Fig. 12.18. The transfer function from θ_o to θ_M is

$$\frac{\Theta_M}{\Theta_o}(s) = \frac{K + J_L s^2}{(D + J_M s)s} \tag{12.6}$$

The frequency response of the linear part of the system (12.6) is shown in vector form in Fig. 12.27. The frequency response of the hysteresis loop is plotted in inverse form to establish the gain and phase shift required of the linear part of the system to bring the total to 180° shift and the loop gain to unity. Fortunately, the curve for the linear portion of the system is independent of amplitude, whereas the curve for the hysteresis loop is independent of frequency. The intersection point A therefore represents the conditions for continuous oscillation. Lowering the factor K lowers the frequency at which the function in (12.6) goes to zero and the angle at which the curve passes through zero. The intersection point, therefore, moves to a lower amplitude of oscillation. Addition of coulomb friction to the analysis could be shown to make the damping constant D in Fig. 12.18 increase for low-amplitude signals. A complete solution including coulomb friction requires the use of families of curves, because the function in (12.6) would then be a function of both frequency and amplitude of signal.

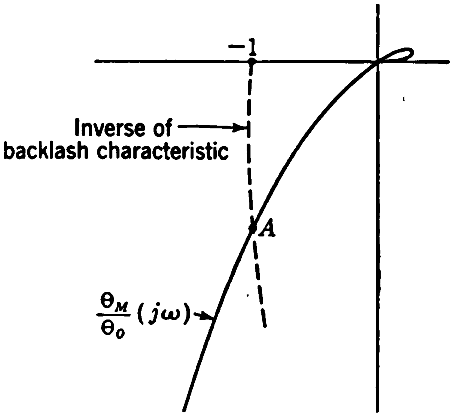


FIG. 12.27. Vector plot of frequency response of system with backlash.

12.10. Conclusions

The phase-plane diagram and the frequency-response method give comparable results for simple systems. The concepts can, therefore, be extended with some confidence to somewhat more complicated systems.

The phase-plane diagram gives an exact solution for a system having no more than two forms of energy storage and capable of being approximated by a series of linear regions. The limitation to two forms of energy storage arises because only two variables can be used in one plot. Another form of energy storage in the system would require that another variable be known at each switching point. This variable would in turn affect the curves for the other two variables, and thus the simplicity of the construction is lost.

The frequency-response method is easily extended to systems with a greater number of energy storages because the linear part of the system is described by a single frequency-response curve. The complexity in this method arises with more than one nonlinear element, because the system as seen by one nonlinear element is then also a function of amplitude of signal.

Another graphical construction is suggested by the fact that the same form of transient is started by each discontinuity in the system. These transients can be traced from a template and added successively for a few cycles.

The various tricks in graphical or analytical treatment have very limited application, however, compared to the total range of the nonlinear problem. They serve more as a means of establishing the effect of such common nonlinearities as relay time delay, coulomb friction, and backlash. Relay action, for example, is a nonlinear form of amplification of a signal and is thus comparable to dead zone in a pick-off or saturation of an amplifier. The time delay in relay action can be treated separately and corresponds to a transfer function of the form e^{-st_1}. Coulomb friction is a nonlinear damping, because the input is velocity and the output is torque or force. Backlash is somewhat unique in that it permits the output to acquire a velocity at each reversal in direction. The effect is destabilizing and is more pronounced for signals whose amplitude is comparable to the backlash itself. The concepts developed from the study of simple systems serve as an aid to judgment for more complicated systems.

The meager conclusions which can be reached for nonlinear systems serve to reemphasize the simplicity of concept in analysis of linear systems. A linear system has characteristics which are independent of the amplitude of the signal, and therefore each signal may be treated independently. All natural transients of the system can be expressed in the form e^{st}. Each transfer function of the system is then a vector which is a function only of the value of s. The natural modes of response of the system are the values of s which make the characteristic determinant equal to zero. After these roots are determined, the amplitudes of individual transient terms can be written immediately. The simplicity of analysis of the linear system is thus based on a series of powerful concepts, which successively reduce the problem to simple steps such as adding angles. Nonlinearities make the characteristics of the system depend upon the total signal present in the system and thus void the basic concepts which make linear analysis simple.

APPENDIX A

THE LAPLACE TRANSFORM

The Laplace transform will be demonstrated by solving a simple problem. Study of the integrals involved will show why one integral is the inverse of the other and give some concept of the transform itself.

Consider the example of a first-degree differential equation

$$x_i = x_o + T\frac{dx_o}{dt} \tag{A.1}$$

in which input $x_i = A_i e^{s_i t}$ and the initial value of the output is $x_o(0)$.

The "transform" of each term in the equation is taken according to the definition

$$X(s) = \mathcal{L}x(t) = \int_0^\infty x(t)e^{-st}\,dt \tag{A.2}$$

For the input term, the transform is

$$\int_0^\infty A_i e^{s_i t} e^{-st}\,dt = A_i \int_0^\infty e^{(s_i-s)t}\,dt = A_i\left[\frac{e^{(s_i-s)t}}{s_i - s}\right]_0^\infty = \frac{A_i}{s - s_i} \tag{A.3}$$

The transform of $x_o(t)$ is by definition $X_o(s)$; a capital letter is always used for a transform. The transform of dx_o/dt requires integration by parts, using the fact that, since

$$d(uv) = u\,dv + v\,du \tag{A.4}$$

then

$$\int u\,dv = \int d(uv) - \int v\,du \tag{A.5}$$

The integral of the derivative term can thus be treated as

$$\int_0^\infty \frac{dx_o}{dt}e^{-st}\,dt = \int_0^\infty e^{-st}\,dx_o = \int_0^\infty d(e^{-st}x_o) - \int_0^\infty x_o(-s)e^{-st}\,dt \tag{A.6}$$

Since the last term is, by definition, $-sX_o(s)$, the right side is evaluated as

$$\int_0^\infty \frac{dx_o}{dt}e^{-st}\,dt = -x_o(0) + sX_o(s) \tag{A.7}$$

The complete expression for the transform of (A.1) thus becomes

$$\frac{A_i}{s - s_i} = X_o(s) + T[-x_o(0) + sX_o(s)] \tag{A.8}$$

Solving for the transform of the output gives

$$X_o(s) = \frac{[A_i/(s - s_i)] + Tx_o(0)}{1 + Ts}$$
$$= \frac{(A_i/T) + x_o(0)(s - s_i)}{(s - s_i)[s + (1/T)]} \tag{A.9}$$

The output function of time is now determined by taking the inverse transform according to the definition

$$x(t) = \mathcal{L}^{-1}X(s)$$
$$= \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} X(s)e^{st}\,ds \tag{A.10}$$

The variable s is a complex number, and the path of integration is a line in the complex plane parallel to the j axis and offset to the right of any pole of the transform.

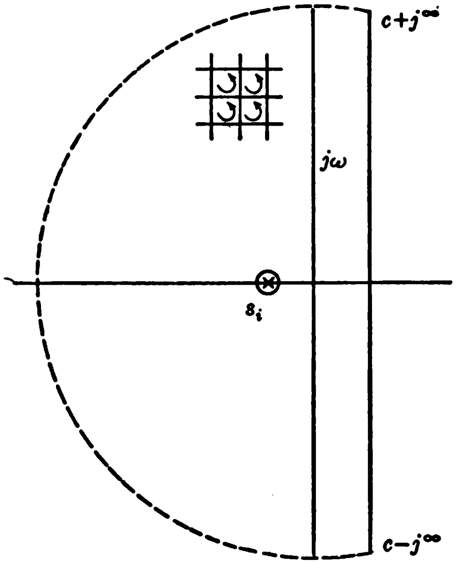


FIG. A.1. Path of integration in the s plane for the inverse Laplace transform.

The inverse transform of $X_o(s)$ can be shown, by use of (A.15), to be

$$x_o(t) = \left[\frac{A_i/T + x_o(0)(s - s_i)}{s + 1/T}\right]_{s=s_i} e^{s_it}$$
$$+ \left[\frac{A_i/T + x_o(0)(s - s_i)}{s - s_i}\right]_{s=-1/T} e^{(-1/T)t} \tag{A.11}$$

Note the form of the result: a term of the form e^{st} appears for each pole of the transform; the amplitude of the term is equal to the transform evaluated at the pole with the factor involving the pole omitted.

The same result is obtained using the methods of Chap. 4 or any other method because the differential equation and the initial conditions are satisfied.

The integral for the inverse Laplace transformation can be shown to produce an opposite effect from the integral for the direct Laplace transformation by taking the simple case of the input itself. The transform of the signal $A_ie^{s_it}$ is shown in (A.3) to be $A_i/(s - s_i)$. The original signal $A_ie^{s_it}$ should be obtained by taking the inverse transform

$$\mathcal{L}^{-1}\frac{A_i}{s - s_i} = \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}\frac{A_i}{s - s_i}e^{st}\,ds = \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}F(s)\,ds \tag{A.12}$$

The significant features of this integration process are shown in Fig. A.1. The path of integration in the complex plane is along the vertical line to the right of the j axis. Completing the path of integration at infinite radius around the left half plane does not add anything to the integral because the factor $1/(s - s_i)$ is zero for this part of the path. The line integral around the entire path can now be considered to be the sum of line integrals around small areas such as those shown by the grid. The value of the line integral around a small area would be zero if the value of the function $F(s)$ in (A.12) were constant. The integral might have a net value, owing to the rate of change of the function $F(s)$ indicated in Fig. A.2, given by

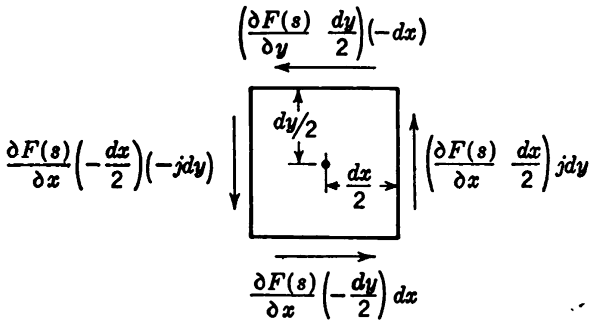

$\left(\frac{\partial F(s)}{\partial y}\frac{dy}{2}\right)(-dx)$
$dy/2$
$\frac{\partial F(s)}{\partial x}\left(-\frac{dx}{2}\right)(-jdy)$
$\frac{dx}{2}$
$\left(\frac{\partial F(s)}{\partial x}\frac{dx}{2}\right)jdy$
$\frac{\partial F(s)}{\partial x}\left(-\frac{dy}{2}\right)dx$

FIG. A.2. Integration around a differential area in the s plane.

$$dA = 2\left[\frac{\partial F(s)}{\partial x}\frac{dx}{2}\right] j\,dy - 2\left[\frac{\partial F(s)}{\partial y}\frac{dy}{2}\right] dx$$

$$= dx\,dy\left[j\frac{\partial F(s)}{\partial x} - \frac{\partial F(s)}{\partial y}\right] \qquad \text{(A.13)}$$

The partial derivatives can be expressed in terms of the derivative

$$\left[j\frac{\partial F(s)}{\partial x} - \frac{\partial F(s)}{\partial y}\right] = \left[j\frac{dF(s)}{ds}\frac{\partial s}{\partial x} - \frac{dF(s)}{ds}\frac{\partial s}{\partial y}\right]$$

$$= \frac{dF(s)}{ds}\left(j\frac{\partial s}{\partial x} - \frac{\partial s}{\partial y}\right) \qquad \text{(A.14)}$$

But $s = x + jy$; therefore $\partial s/\partial x = 1$ and $\partial s/\partial y = j$, making the value of the final parenthesis equal to $j - j$, or zero.

If the derivative $dF(s)/ds$ has a finite value, then the integral around any small area dA is zero. The line integral around the complete left half plane has a value then determined only by the area containing the pole s_i at which the derivative is infinite. The value of the integral enclosing this pole can be obtained readily by choosing a circular path about the pole as a center, as shown in Fig. A.3.

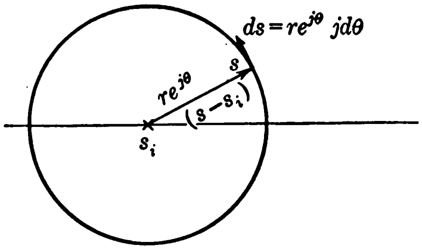


FIG. A.3. Integration around a pole in the s plane.

The integral in (A.12) becomes

$$\frac{1}{2\pi j}\oint \frac{A_i e^{st}}{s-s_i}\,ds = \frac{A_i}{2\pi j}\int_0^{2\pi} \frac{e^{s_it}}{re^{j\theta}}\,re^{j\theta}j\,d\theta = \frac{A_ie^{s_it}}{2\pi j}\,2\pi j \tag{A.15}$$

The value of the inverse transform is thus simple $A_ie^{s_it}$, as originally specified in (A.3).

If the transform has several poles, then the integral involved in the inverse transform encloses all these poles. The integral is the sum of terms of the form $A_me^{s_mt}$, in which s_m is a pole and A_m is the value of the transform with $s = s_m$ and the factor $s - s_m$ omitted. This explanation is intended primarily to give a basis for understanding the Laplace transform. The concepts involving line integral in the complex plane are rather far removed from the problem of describing a control system, however, and thus the topic of Laplace transform is given here as an appendix.

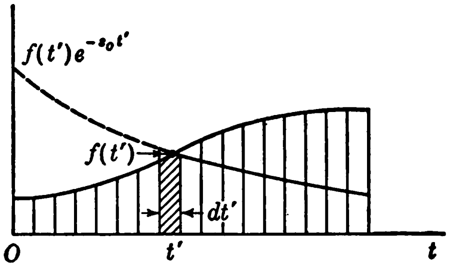


Fig. A.4. Graphical picture of convolution.

An interesting concept of the significance of a transform can be obtained by considering another simple system:

$$f_i(t) = -s_ox_o + \frac{dx_o}{dt} \tag{A.16}$$

The input $f_i(t)$ is arbitrary in form and is considered to be a series of impulses, as shown in Fig. A.4.

A unit-impulse input would cause a transient output e^{s_ot}. The impulse $f(t')\,dt'$ at $t = t'$ will cause a transient after $t = t'$ of

$$dx_o = f(t')\,dt'\,e^{s_o(t-t')} = [f(t')\,dt'\,e^{-s_ot'}]e^{s_ot} \tag{A.17}$$

The transient has only the decay time $t - t'$ between the time of the impulse t' and the time t being considered. A larger impulse $f(t')\,dt'e^{-s_ot'}$ acting at the initial time $t = 0$ would produce the same transient at time t.

The total output at time t can be found as the integral of the effect of each of the impulses of the input:

$$x_o = \int_0^t f(t')\,dt'\,e^{s_o(t-t')} = e^{s_ot}\int_0^t f(t')e^{-s_ot'}\,dt' \tag{A.18}$$

The factor e^{s_ot} is a constant for the given integral and thus may be removed as a factor. The remaining integral has the same form as the definition of the Laplace transform, except that the upper limit of integration is t rather than infinity. The integral represents the impulse acting at time $t = 0$ which would produce the same value of the output at the time t as does the actual input. If the input is of the form $A_1e^{s_it'}$, then the integral can be found and evaluated at the limits 0 and t.

$$
\begin{aligned}
x_o &= e^{s_o t} \int_0^t A_1 e^{s_1 t'} e^{-s_o t'}\, dt' = e^{s_o t} A_1 \left[\frac{e^{(s_1 - s_o)t'}}{s_1 - s_0} \right]_{t'=0}^{t'=t} \qquad \text{(A.19)} \\
&= e^{s_o t} A_1 \frac{1}{s_o - s_1} + e^{s_1 t} A_1 \frac{1}{s_1 - s_o}
\end{aligned}
$$

Note that the integral evaluated at the limit $t = 0$ gives the amplitude of the term involving $e^{s_o t}$. The value of the integral in (A.18) evaluated only at the limit $t = 0$ then gives the effective value of the impulse acting at $t = 0$ which would produce the same amplitude of the $e^{s_o t}$ term as the actual input. The Laplace transform of the input evaluated for $s = s_o$ is then exactly equal to the effective impulse determined above, because the value of the integral in (A.18) evaluated at $t = \infty$ is always zero.

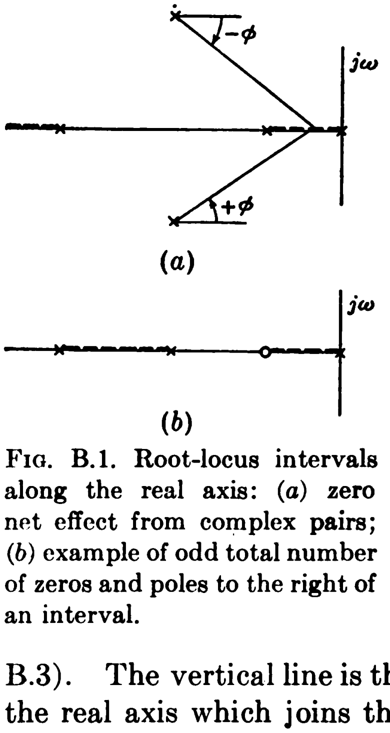
A standard method of explaining the Laplace transform starts from the Fourier series representation of a periodic function of time. The period is then allowed to approach infinity, which causes the Fourier series to become the Fourier integral. A decay factor e^{-st} is then introduced in order to make the integral convergent, making the Fourier integral become the direct Laplace transform integral. The inverse process of expressing a function of time from a Fourier series in turn becomes the inverse Laplace transform integral. For a rigorous and complete treatment of the Laplace transform, see Gardner and Barnes, "Transients in Linear Systems."

APPENDIX B

AIDS IN SKETCHING ROOT-LOCUS PLOTS

The main aids in sketching root-locus plots are the many special cases, such as intervals along the real axis and the asymptotes for large values of s. The intermediate portions of the plot can then usually be sketched by seeking to keep the net change in angle zero or by using the charged-particle analogy.

The following aids for sketching root-locus plots are used throughout the book in solving the various problems; this appendix is intended as a summary. The simplest portions of the plot to establish are the intervals along the real axis, because all angles are then either 0° or 180° (Fig. B.1). Complex pairs of zeros or poles contribute no net angle for points along the real axis. For a 180° locus, the locus will exist at intervals along the axis which have an odd total number of zeros and poles to the right of the interval.



(*a*)

(*b*)

FIG. B.1. Root-locus intervals along the real axis: (*a*) zero net effect from complex pairs; (*b*) example of odd total number of zeros and poles to the right of an interval.

For very large values of s, all angles are essentially equal (Fig. B.2). The locus will thus finally approach asymptotes[1] at the angles

$$\frac{180^\circ \pm n360^\circ}{\text{Number of poles} - \text{number of zeros}} \tag{B.1}$$

A quadratic system provides patterns which often exist in essentially the same form in more complicated systems (Fig. B.3). The vertical line is the perpendicular bisector of the interval along the real axis which joins the two poles in Fig. B.3*a*. The locus in Fig. B.3*b* is a circle.

Breakaway points from the real axis occur where the net change in angle caused by a small vertical displacement is zero. In Fig. B.4*a*, the point P satisfies this condition if $1/b_0 = 1/b_1 + 1/b_2$. The contribution of a complex pair of poles to the phase angle just above the real axis can be estimated by establishing equivalent points along the real axis, as shown in Fig. B.4*b*. The phase angle β subtended at A is equal to $(OP/AP) \sin \phi$. The same angle β is subtended at B if $OB = AP/\sin \phi$. The point B can be located by sketching a line from A perpendicular to the line OA.

---

[1] The use of the sum of the roots as the center of the asymptotes to achieve greater accuracy was developed by Robert M. Stewart, of California Institute of Technology.

The initial direction at which a locus emerges from a pole can be established by considering trial values of s on a small circle around the

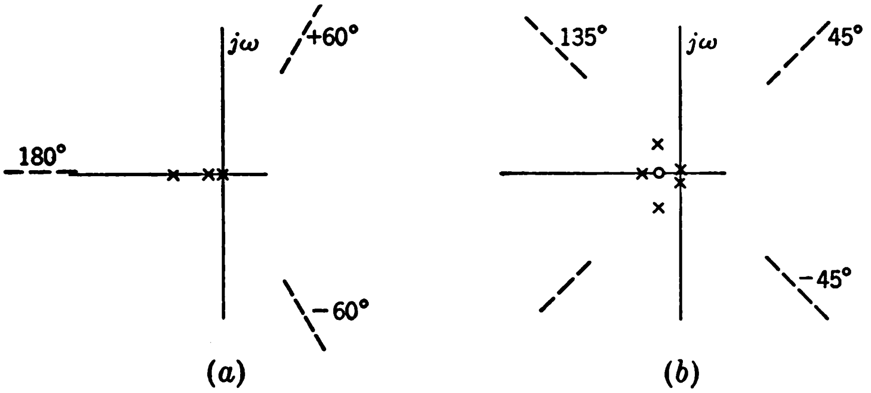


FIG. B.2. Final asymptotes for root loci: (*a*) 60° asymptotes for system having three poles; (*b*) 45° asymptotes for system having an excess of four poles over zeros.

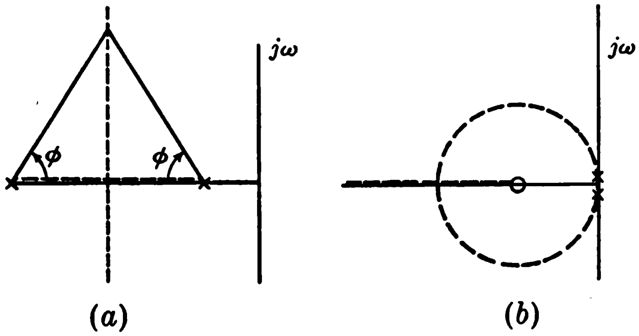
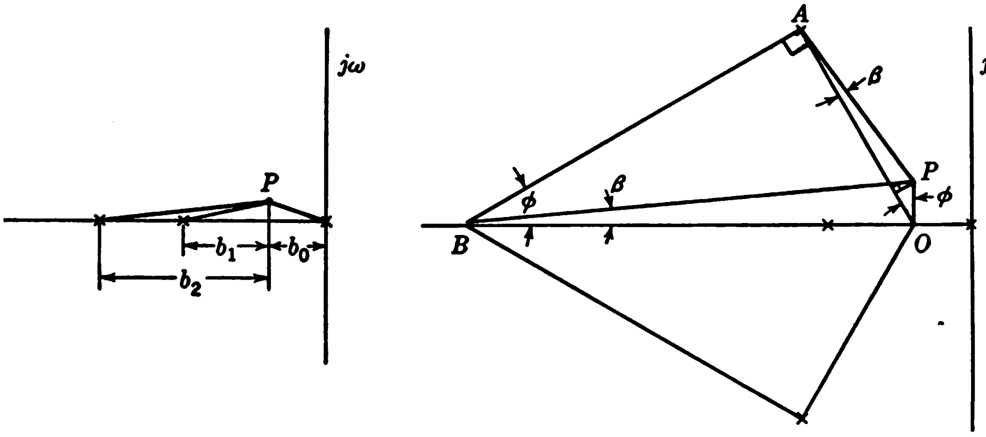


FIG. B.3. Quadratic-system plots: (*a*) perpendicular bisector; (*b*) circle for system having one zero and two equal poles.

pole. The angle of the vector from the pole to the trial value of s is selected so as to make the total angle 180°. One method, shown in Fig. B.5*a*, is to start with a reference line to the left of the point and then reduce the angle of the vector for each angle contributed by some other vector. Note that the angles need merely be sketched; numerical addition is not necessary.

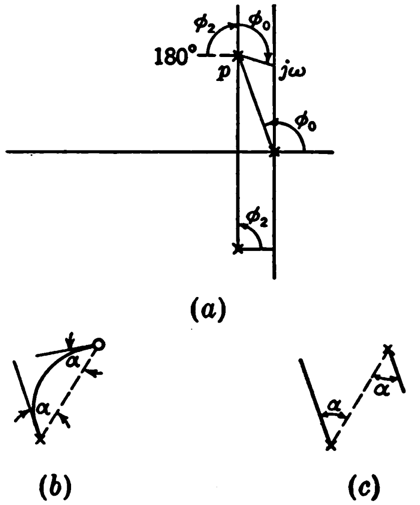
Figure B.5*b* illustrates the case of a pole and zero which are reasonably far removed from any other zeros or poles in the plot. If the initial direction of the locus from the pole has been determined, then the direction of the locus at the zero can be found by drawing a line between the pole and zero and making angle α be the same. In most cases the two points join in a circular-shaped arc. Figure B.5c illustrates a similar case in which the two points are both poles. The loci then emerge parallel to each other but opposite in direction.



(a) (b)

FIG. B.4. Breakaway points: (a) net sum of reciprocals of base lengths is zero; (b) complex pair of points replaced by equivalent pair of points along the real axis.

The analogy between a root-locus line and a flux line in an electric field plot is useful in sketching the remainder of a locus once the starts have been determined. Thus a locus follows the path which a charged particle would follow if repelled from the poles and attracted to the zeros with a force inversely proportional to the distance. The locus thus emerges from the pair of poles at the origin of Fig. B.6 and starts to curve around toward the zero, but the effect of the quadratic pair of poles is then to bend the locus back to the right.



(a)

(b) (c)

FIG. B.5. Initial directions of loci: (a) subtraction at a pole of angles due to effect of other poles; (b) locus for an isolated pole and zero; (c) loci for an isolated pair of poles.

Figure B.7 illustrates two useful facts explained further in Art. 8.4: the fixed sum of roots for systems having two or more poles than zeros, and the substitution of roots for poles as a means of continuing a plot. The familiar locus for the cubic system is carried out to the point where the complex roots cross the j axis. The distance by which the pair of roots have moved to the right compared to the original position of the poles is then equal to the distance by which the real root has moved to the left of its pole. Moreover, a plot which is

Digitized by Google Original from UNIVERSITY OF MICHIGAN

based on the roots, shown as triangles, would continue along the same path as the original locus. The use of the roots, however, permits the direction of the locus to be established as ϕ° off the vertical.

Figure B.8 shows systems which are nearly identical, although the first impression created by their root-locus plots is that they are quite different. Figure B.8b shows the critical case, in which the roots from the complex poles move straight down and meet the roots from the real axis. In Fig. B.8a, the poles are shifted slightly to the left, and the loci do not meet, as shown. Similarly, in Fig. B.8c, the poles are slightly to the right, and the loci now curve away in directions opposite to Fig. B.8a.

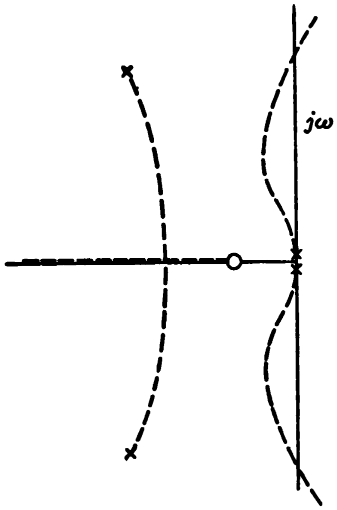


FIG. B.6. Sketch of locus using the charged-particle analogy.

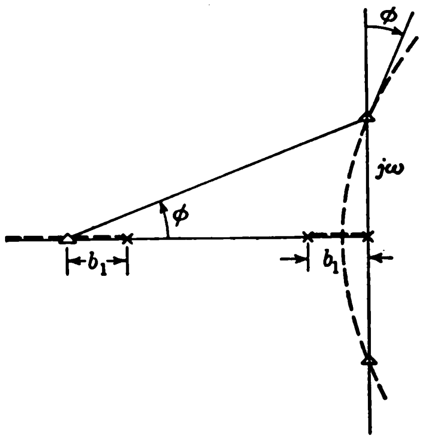
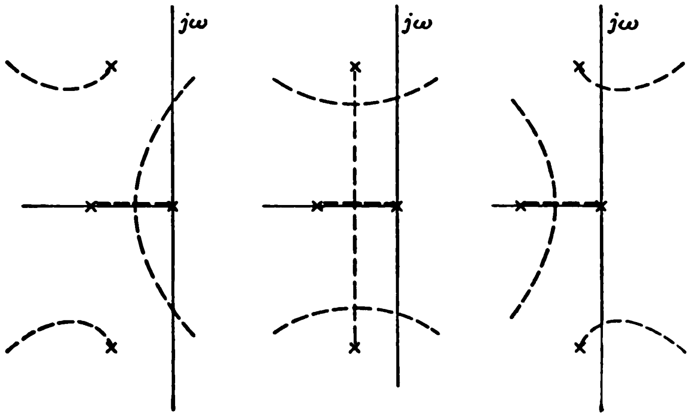


FIG. B.7. Example illustrating the fixed sum of roots and the replacement of poles by roots.



(a) (b) (c)

FIG. B.8. Plots for equal or nearly equal roots: (a) plot for complex poles slightly to the left of critical position; (b) critical plot for equal roots; (c) plot for complex poles slightly to the right.

The root-locus plots in this case exaggerate the difference between the systems; if a transient response or frequency response were determined for the three cases, they would be found to be nearly the same. The region near the junction point of the loci in Fig. B.8*b* is indefinite; any point in the region will give nearly 180° shift and essentially the same value of gain. This condition exists near any point of equal roots, as can be shown by solving for the rate of change of the function below with respect to *s*:

$$\left[\frac{d}{ds}(1 + Ts)^2\right]_{s=-1/T} = [2(1 + Ts)T]_{s=-1/T} = 0 \qquad \text{(B.2)}$$

At the equal-root point, the rate of change of the function is zero, and for points near this point the rate of change is small. The effect is even more pronounced with three or more equal roots.

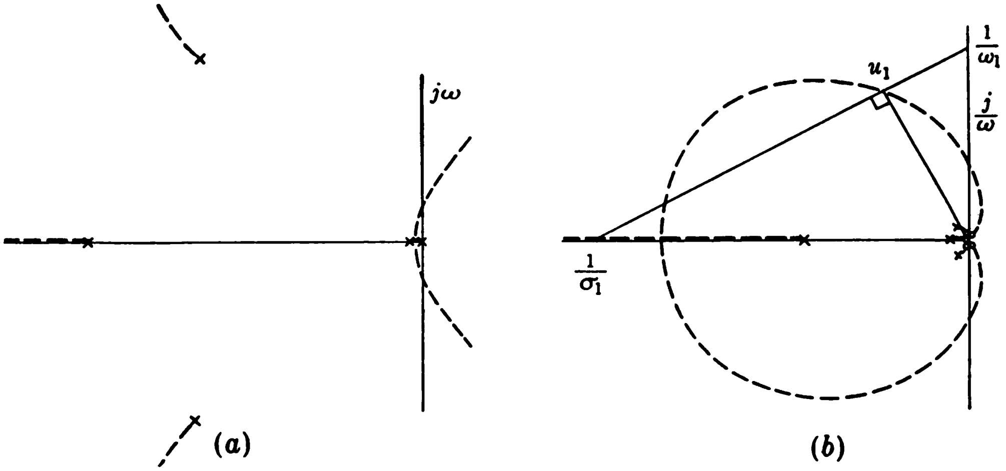


FIG. B.9. Reciprocal plot obtained by substituting $s = 1/u$ in the characteristic equation: (*a*) standard plot in the *s* plane; (*b*) plot in the *u* plane.

Occasionally the roots near the origin may be of particular interest, whereas the high-frequency effects are comparatively unimportant. A root-locus plot can be turned "inside out" by making a change in variable such as $s = 1/u$. Thus the original *s* plot in Fig. B.9*a* shows a low-frequency oscillation in which the poles far removed from the origin have a small effect. Substitution of $s = 1/u$ into the original equation gives a new equation in *u* whose plot is shown in Fig. B.9*b*. The poles which are distant from the origin in Fig. B.9*a* are close to the origin in Fig. B.9*b*, although the angles are the same. Similarly all points on the locus have the same angle as in the first plot; the magnitudes are merely the reciprocals of their previous values. The new plot has zeros at the origin corresponding to the excess of poles over zeros in the original plot. A point u_1 corresponds to a root s_1. The values of $1/\sigma_1$ and $1/\omega_1$ are obtained directly on the *u* plot by drawing a perpendicular to the radial line to that point and noting the intersection with the real and j axes. The plot is, in general, more difficult to make because of the added zeros and because the original s plot is more familiar. The plot does, however, have the advantage of permitting extra high-frequency poles to be added almost indefinitely in studying their effect upon the low-frequency roots. For the purpose of approximating phase angles, a cluster of pairs of zeros and poles near the origin can be replaced by a single pair having the same total horizontal separation. A similar simplification could be gained on the original s plot, however, by replacing quadratic points by points along the real axis, as in the determination of breakaway points, and then finding a single point on the real axis which contributes the same phase shift near the origin as the combination of points farther out.

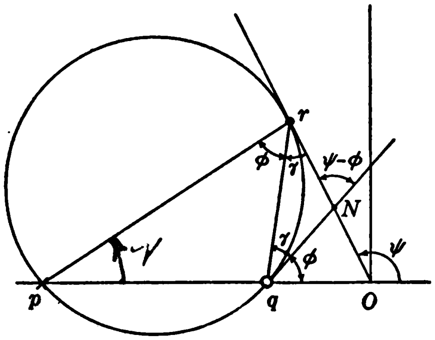


FIG. B.10. Geometry of a plot having a given phase lead from a circuit having minimum attenuation of a d-c signal.

In some cases the desired position of a root is known. The total phase angle due to the given part of the system can be measured at this root, and it is necessary to determine a lead network so that the whole system will have the desired root. Let ϕ be the phase lead necessary to make the net phase angle equal to 180°. The locus of points in the s plane for which a network gives a lead angle ϕ is a circle, as shown in Fig. B.10.

For the lead network to have minimum attenuation of a d-c signal, the circle should be tangent to the line of constant damping ratio at the desired root. A tangent to the circle at the zero q just makes the angle ϕ with respect to the real axis. The two tangents intersect at the point N. The tangents Nr and Nq are equal, and therefore the triangle Nqr is isosceles with the base angles γ equal. The angle between the two tangents at N is the difference $\psi - \phi$, as shown. Both 2γ and $\psi - \phi$ are the supplements of the angle rNq and therefore are equal to each other.

$$2\gamma = \psi - \phi \qquad \text{or} \qquad \gamma = \tfrac{1}{2}(\psi - \phi) \tag{B.3}$$

In the given problem, therefore, the line rq is drawn at the angle γ with respect to the line Nr in order to locate the zero q. The line rp is drawn through r at the angle ϕ with respect to the line rq in order to locate the pole *p*. The attenuation ratio of the network is then measured as the ratio of line lengths Oq/Op.[1]

If the required lead angle is greater than can be achieved with a single-stage network, then a double or perhaps a triple lead network should be used. The angle ϕ in Fig. B.10 then becomes the lead angle required per stage. For minimum attenuation of d-c signals, the phase lead per stage should be small.

---

[1] The method of selecting the desired position of a root and then solving analytically for the position of the zero and pole of a lead network was described in a North American Aviation memo by W. E. Dickinson, now with International Business Machines at San Jose, Calif.

APPENDIX C

USE OF THE SPIRULE[1]

The Spirule[2] is a plastic device which permits rapid addition of angles or multiplication of lengths of vectors. The spirule consists of an arm and disk held together with a light friction fit by a special eyelet, which also serves as a pivot point. On a root-locus plot, the pivot point is placed at a trial s point and the arm rotated, with respect to the disk, through each of the vector angles, to obtain their sum. A logarithmic spiral curve on the arm permits the logarithm of a vector length to be obtained as an angle, so that the addition of such angles corresponds to the addition of logarithms. The Spirule can also be used in plotting vectors or as a circular slide rule.

The Spirule is most easily explained by using it with a simple example, such as checking the root r_1 in Fig. 7.15, which is duplicated for convenience in Fig. C.1. The pivot point is placed at a trial position, r_1, on the roughly sketched locus of roots. The small pin serves to hold it there. The zero angle of the disk is set at the radial edge of the arm; this edge is brought in line with the pole at the origin, as shown in Fig. C.2*a*.

The spiral curve and the logarithm scale on the disk are not involved in this operation of adding phase angles. The pivot point can be held at the r_1 point by lightly pressing on the Spirule at the point with the right thumb. The index finger of the right hand is then in position to press the disk against the paper to keep it from moving while the arm is rotated with the left hand until the radial edge is horizontal, as shown in Fig. C.2*b*.

Note that the arm is thus rotated through the angle ϕ_0 *with respect to the disk;* the angle may be read on the disk at the radial edge of the arm as being about 240°, or $-120°$, the minus sign due to the angle being a lag angle. The index finger is then removed from the disk while the arm is rotated counterclockwise until the radial edge is in line with the pole at $-1/T_1$. Note that the friction between the arm and the disk must be sufficient to make the disk move with the arm, because the dial reading should not change in this motion.

---

[1] The Spirule can be obtained from the Spirule Company, 9728 El Venado, Whittier, Calif.

[2] The idea of building the principles of the Spirule into a plastic device was conceived by J. F. Schmidt, and the first model was built by him. The name "Spirule" was coined by W. D. Lyon as a contraction of "spiral" and "slide rule." Both men are engineers in the Electromechanical Division, North American Aviation, Inc., Downey, Calif.

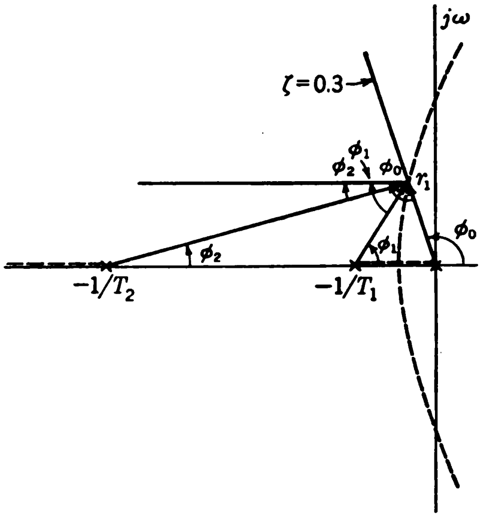
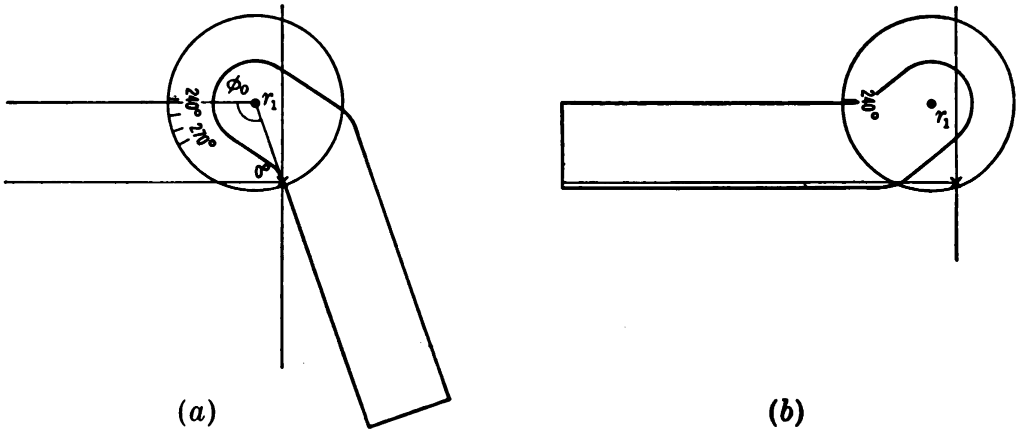


FIG. C.1. Typical root-locus plot.

For adding the angle ϕ_1, the index finger again is used to press the disk against the paper while the arm is rotated until the radial edge is horizontal. The procedure is repeated for the last angle ϕ_2. The edge of the arm is now approximately opposite its starting position with respect to the disk. The sum of the phase angles is read on the disk at the edge of the arm. The difference between this sum and the desired sum of 180° can be marked on the plot. A second trial point should be selected slightly to the left of the point r_1 because all the phase angles are thus increased. If this point is still in error, the true point should be located by interpolation between the two trial points.

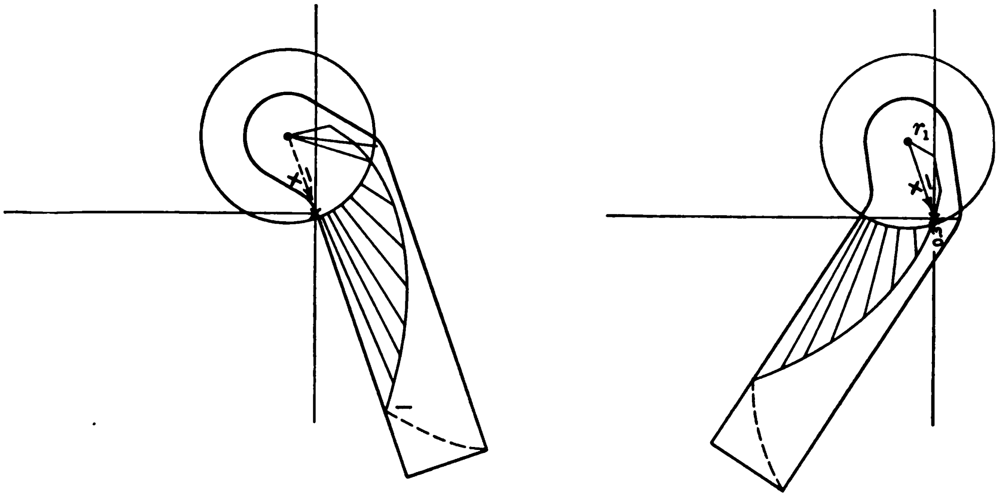


(*a*) (*b*)

FIG. C.2. Use of Spirule for adding angles: (*a*) initial position; (*b*) after rotation through ϕ_0.

All the vector lengths which vary with s have one end at the s point. The product of these lengths is proportional to the gain needed to make that s point a root, and such a product is conveniently obtained by adding the logarithms of the lengths. The Spirule has provision for adding angles; it could serve also to obtain the product of the lengths if a conversion from a vector length to a logarithm in the form of an angle were provided. This conversion is achieved by use of the logarithmic spiral curve, as will be demonstrated for the plot in Fig. C.1.

The scale of the logarithmic spiral curve on the Spirule is such that unity, the distance from the pivot point to the point where the curve intersects the radial edge, is 5 in. The radial distance to this scale from the pivot point to the curve is marked on the curve as shown. The angle between the radial edge and the line from the pivot point to a point on the curve is proportional to the logarithm of the radial distance to that point on the curve. Note that the logarithm of 0.1 is −1, which is represented by an angle of 90°.



(*a*) (*b*)

FIG. C.3. Use of Spirule for multiplying vector lengths: (*a*) initial position; (*b*) after rotation through 90° $\log_{10} |r_1|$.

The Spirule, pivoted at r_1, is initially set with the edge of the arm in line with the first pole and the arrow on the disk in line with the edge, as shown in Fig. C.3*a*. The disk is held with the index finger while the arm is rotated until the *spiral curve meets the pole*, as shown in Fig. C.3*b*.

The length of this vector is read, on the logarithmic spiral scale on the arm at the arrow marked on the disk, as 0.30. The scale of the plot is changed in printing this figure, and therefore a check cannot be made by using a Spirule with this plot. The angle through which the arm rotated with respect to the disk is the logarithm of this length to the scale factor for angle given by

$$\frac{\psi_0°}{90°} = \frac{\log_{10} |r_1|}{\log_{10} 0.1} \tag{C.1}$$

The disk is now released while the arm is rotated until the edge is in line with the pole at $-1/T_1$. Note that the disk moves with the arm during this motion, and the reading does not change. The disk is held with the index finger, however, while the arm is rotated until the spiral curve meets the pole at $-1/T_1$. The arm is thus rotated with respect to the disk through the angle ψ_1, which is the logarithm of the vector length $|1/T_1 + s|$. The arm has now rotated with respect to the disk through the sum of the angles ψ_0 and ψ_1. The operation is repeated for the pole $-1/T_2$.

This value on the scale is, then, the product of the vector lengths in the *numerator* of the equation for gain K given by

$$K = \frac{|s||1/T_1 + s||1/T_2 + s|}{|1/T_1||1/T_2|} \tag{C.2}$$

The product of lengths in the denominator can be found by using the Spirule in a similar fashion, using the origin as the pivot point and considering length to the poles at $-1/T_1$ and $-1/T_2$. The quotient, giving the value of K, may be found as a separate slide-rule operation or by using the Spirule as discussed at the end of this article. Note that any difference in scale factor between the plot and the Spirule would not be important for ratios; it does have to be considered for the vector s itself. Here one should multiply the previously obtained value of gain by the *numerical value* on the plot corresponding to unity on the Spirule, or 5 in.

In solving any problem, the gain for several points along the locus is usually desired. Hence it is convenient to note all the corrections for base values and scale factor as a single angle of rotation of the disk with respect to the arm. This angle can be marked on the disk and reset before making the calculations at any s point. The gain K can then be read directly on the scale after all rotations with the Spirule at the s point.

If the transfer function contains a zero as well as poles, the logarithm of the vector length to the zero should be *subtracted* rather than added. Subtraction can be accomplished conveniently by originally lining up the zero *on the spiral curve* and rotating the arm until the radial edge meets the zero.

If a length is greater than 1, the spiral curve would normally go off the plastic arm, but the curve is folded back as shown by the dotted curve. This folding is done in such a way that the angle between the curve and the radial edge is still the logarithm of the radial length. The direction of rotation should be reversed, however, having the pole move from the curve to the edge. This rotation corresponds to the fact that the logarithm of a number greater than 1 is positive, whereas the logarithm of a number less than 1 is negative.

The accuracy in using the Spirule becomes less for points close to the pivot point, which is a natural characteristic of the plot itself. Again it is well to note that any distorted scale of the plot itself would not permit the basic operation of adding two vectors, such as $1/T$ and s to obtain $1/T + s$, which is essential to the given use of the plot.

The Spirule can be used as a slide rule with the use of a single straight line, rather than a root-locus plot. To multiply two numbers, the disk is set with the $X1$ arrow of the disk aligned with the radial edge of the arm and the pivot point of the Spirule placed at any point at least 5 in. to the right of the end of the line. The arm is then rotated with the disk held fixed until the spiral curve on the arm intersects the line at the numerical value of the first number. The disk is then released while the edge of the arm is again aligned with the straight line. The disk is held fixed again while the arm is rotated until the curve intersects the line at the second number. The product is read on the arm in line with an arrow of the disk. To divide by a number, the rotation of the arm, with respect to the disk, should be in the opposite direction.

APPENDIX D

ROUTH'S CRITERION

Routh's criterion determines the number of roots having a positive real part by simple calculations using the coefficients of the characteristic equation. An equivalent circuit provides direct analogies and a short proof.

Routh's criterion is perhaps the oldest and, for some problems, still the best method of determining the stability of a system. It requires only simple calculations starting from the coefficients of the characteristic equation, as will be demonstrated for the case of a fourth-degree system.

$$a_4s^4 + a_3s^3 + a_2s^2 + a_1s + a_0 = 0 \tag{D.1}$$

The coefficients of the even powers of s are placed in the first row and those of the odd powers of s in the second row.

$$\begin{array}{ccc} a_4 & a_2 & a_0 \\ a_3 & a_1 & \end{array}$$

Elements of a third row are calculated from the above two rows according to the pattern

$$b_2 = \frac{a_3a_2 - a_1a_4}{a_3} \qquad b_0 = \frac{a_3a_0}{a_3} \tag{D.2}$$

Similarly, elements for successive rows are formed until a complete triangular array of numbers is formed:

$$\begin{array}{ccc} a_4 & a_2 & a_0 \\ a_3 & a_1 & \\ b_2 & b_0 & \\ c_1 & & \\ d_0 & & \end{array} \tag{D.3}$$

where $c_1 = (b_2a_1 - b_0a_3)/b_2$ and $d_0 = c_1b_0/c_1$.

The number of roots having positive real parts can be shown to be equal to the number of sign changes occurring in the left-hand column of the array.

The simplicity of the pattern suggests that a correspondingly simple proof should exist. The key idea is to set up an equivalent form of the characteristic equation which permits successive simplifications.

$$\frac{a_4s^4 + a_2s^2 + a_0}{a_3s^3 + a_1s} + 1 = Z(s) + 1 = 0 \tag{D.4}$$

This form of the equation can represent the voltage drop around a circuit containing a resistance of 1 ohm and an impedance $Z(s)$. The impedance $Z(s)$ can be simplified by dividing numerator by denominator, giving

$$Z(s) = \frac{a_4}{a_3}s + \frac{[a_2 - (a_1a_4/a_3)]s^2 + a_0}{a_3s^3 + a_1s} = L_1s + Z_1(s) \tag{D.5}$$

The first term corresponds to an inductance. Note that the coefficients in the numerator of the second term correspond to the elements in the third row of the array (D.3) used in Routh's criterion; that is,

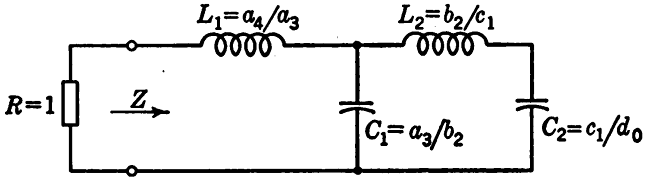


FIG. D.1. Equivalent circuit for a system established by Routh's criterion calculations.

$$a_2 - \frac{a_1a_4}{a_3} = b_2$$

and $a_0 = b_0$.

A similar process of division can be carried out if the reciprocal of $Z_1(s)$, the admittance $Y_1(s)$, is considered.

$$Y_1(s) = \frac{1}{Z_1(s)} = \frac{a_3s^3 + a_1s}{b_2s^2 + b_0} = \frac{a_3}{b_2}s + \frac{[a_1 - (a_3b_0/b_2)]s}{b_2s^2 + b_0} = C_1s + Y_2(s) \tag{D.6}$$

The first term C_1s is now a capacitance, and the coefficients in the numerator of the second term now correspond to the element $c_1 = a_1 - b_0a_3/b_2$ in the fourth row of Routh's array. Considering the reciprocal of $Y_2(s)$ in turn gives

$$Z_2(s) = \frac{1}{Y_2(s)} = \frac{b_2s^2 + b_0}{c_1s} = \frac{b_2}{c_1}s + \frac{b_0}{c_1s} = L_2s + \frac{1}{C_2s} \tag{D.7}$$

The impedance $Z(s)$ is now completely determined as a ladder network, shown in Fig. D.1.

Note that each element of the circuit has a numerical value which is the ratio of the successive elements in the left column of Routh's array. A circuit element is therefore negative for each sign change which appears in this column. The pattern of calculation of the array thus corresponds directly to division of one function by another, and the elements in the left-hand column take on definite significance in terms of circuit elements.

The remainder of the proof is to show that the number of roots having a positive real part is equal to the number of circuit elements having a negative sign. A circuit element with a negative sign is unusual in that the voltage across the element will tend to aid rather than oppose a current flow. A positive capacitance, for example, builds up a voltage proportional to the integral of current into the capacitor in a sense such as to oppose further current flow in the same direction. A negative capacitance, on the other hand, would have the corresponding voltage aid further current flow. If one circuit element changes from positive to negative, then one root must shift from the left half plane to the right half plane for Routh's criterion to be valid. To study the special case when C_2 in Fig. D.1 changes sign, consider the rest of the circuit as seen from the terminals of C_2 to be an impedance Z_{C2}. This impedance will have zeros and poles, but its value for d-c signals is simply R; therefore, the impedance can be expressed as

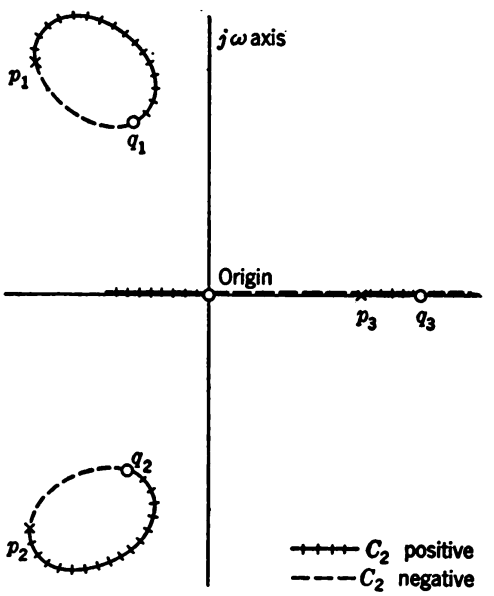


FIG. D.2. Typical plot illustrating a root crossing the j axis when the sign of a capacitor changes.

$$Z_{C2} = \frac{R(1 - s/q_1)(1 - s/q_2) \cdots}{(1 - s/p_1)(1 - s/p_2) \cdots} \tag{D.8}$$

The equation for roots of the complete circuit can now be expressed as

$$-1 = Z_{C2}C_2s = \frac{RC_2s(1 - s/q_1)(1 - s/q_2) \cdots}{(1 - s/p_1)(1 - s/p_2) \cdots} \tag{D.9}$$

If the absolute value of C_2 is very large, one root must be near the origin and the other roots near the zeros of Z_{C2}.

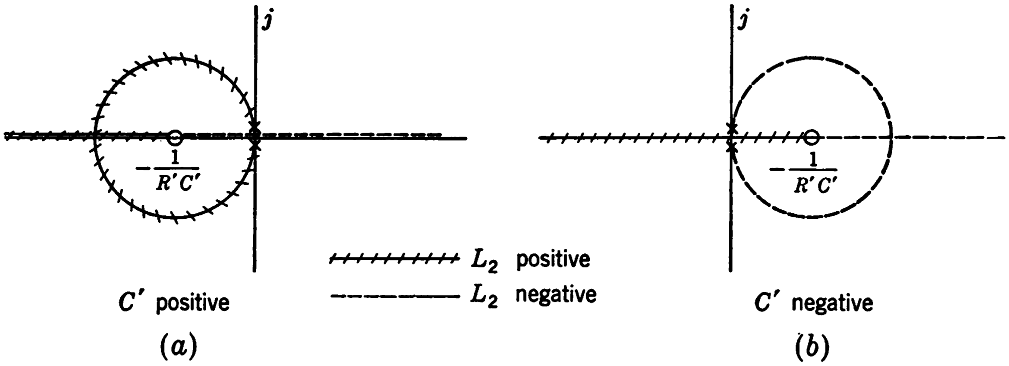
For s near the origin, however, the impedance Z_{C2} is essentially just R; therefore, the locus of roots simply passes through the origin when C_2 changes from $+\infty$ to $-\infty$. The other roots remain near the zeros of $Z_{C2}(s)$ when C_2 changes sign at infinity (Fig. D.2).

The effect of changing the sign of an inductance can be studied in similar fashion. The impedance of the circuit seen from an inductance such as L_2 is of the form $R' + 1/C's$ for values of s near zero. The equation for roots near the origin is thus

$$\frac{R'C's + 1}{C's}\frac{1}{L_2 s} = -1 \tag{D.10}$$

The two possible plots using all combinations of signs for C' and L_2 are shown in Fig. D.3. Note that one more root has a positive real part if the sign of $L2$ is negative rather than positive.

The above special cases show that the number of roots in the right half plane is increased by 1 if any circuit element changes from positive to negative. Some additional arguments are needed to show that these are the only conditions for a root to shift to the right half plane. First, with all circuit elements positive, the system is stable because no energy source is present in the system, whereas one resistance is present to dissipate energy. This argument does not depend upon the magnitude of any inductance or capacitance. In a general case some of the circuit elements are negative, and thus some of the roots are in the right half plane. It remains to be shown that none of the roots will cross the j axis for any magnitude of an inductance or capacitance with a given sign. The argument here is that, for a root to cross the j axis, the impedance as seen from the capacitance or inductance must be purely reactive. This impedance will always have a real part, however, owing to the one resistance which cannot be canceled by any other element in the circuit. A negative resistance would be necessary, but the manner in which the equivalent circuit is set up permits only the single resistive element to be present.



C' positive (a)

C' negative (b)

FIG. D.3. Examples of shift of roots across the j axis when the sign of an inductance is changed.

The picture presented by an equivalent circuit shows more information than is usually stated in Routh's criterion itself. The circuit shows the manner in which the roots cross from the left to the right half plane. A capacitance going through infinity, for example, corresponds to a real root passing through the origin as in Fig. D.2. Other possibilities not previously considered are roots at infinity corresponding to values of C near zero or values of L near infinity. The special cases of C or L being zero or infinite correspond to open circuits or short circuits in branches of the equivalent circuit, so that a simpler solution is possible for the other roots of the system. All the results given here in terms of an equivalent electrical circuit could, of course, be restated in terms of an equivalent mechanical system if desired.

An extension of Routh's criterion determines the position of the roots with respect to some other line in the complex plane parallel to the j axis. Thus, if the variable $\sigma_1 + s'$ is substituted for the original variable s, the $j\omega$ axis of the s' plane corresponds to the $\sigma_1 + j\omega$ axis of the original plane. Routh's criterion then determines the numbers of roots having less than the specified damping for σ, a negative number.

APPENDIX E

ANALOGY BETWEEN FLUX PLOTS AND ROOT-LOCUS PLOTS

A complete root-locus plot is a grid of curvilinear squares similar to that formed by an electrostatic field. The analogy is useful in sketching a root-locus plot and permits a machine solution.

The concepts associated with electrostatic flux plots and root-locus plots are completely different; yet each can be shown to involve the same pattern of curvilinear squares (Fig. E.1). The dotted lines are always normal to the solid lines. If any curvilinear square were sufficiently subdivided, the basic units would be true squares.

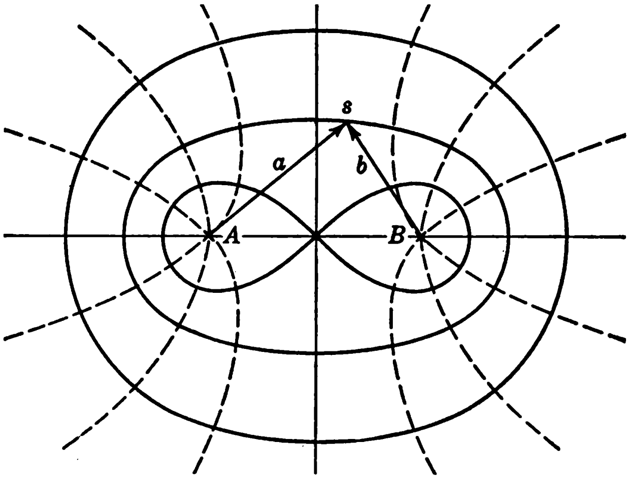


FIG. E.1. Curvilinear square grid for a system having two poles.

For the electric field the solid lines are lines of equipotential, and the dotted lines are lines of flux. The field is formed by two long wires; perpendicular to the plane of the figure, at A and B, equally charged to the same density per unit length. The field intensity at point s due to the wire at A varies inversely with the distance a; the total work to bring a unit charge from large distance to point s varies as $\log a$. The potential at s due to the charged wire at B varies as $\log b$; therefore, the total potential at point s is

$$\text{Potential} = \text{constant} + \log a + \log b \tag{E.1}$$

The lines of flux indicate the direction of the field intensity; therefore they cross lines of constant potential at right angles. The strength of the field varies as the density of the flux lines; doubling the strength of the field would halve the spacing between flux lines and would halve the distance between lines of constant potential. The electric field thus fulfills all the requirements of a grid of curvilinear squares.

For the root-locus plot, the solid lines will be shown to correspond to constant magnitude of the function $F(s)$, and the dotted lines to constant phase angle. Consider first the logarithm of a general function:

$$\log F(s) = \log Re^{j\phi} = \log R + j\phi \tag{E.2}$$

The locus of roots is the path of values of s which hold ϕ constant at π radians while $\log R$ varies. Another path in the s plane exists which will keep $\log R$ constant while ϕ varies. The plot in Fig. E.1 corresponds to the function $F(s) = |a|e^{j\phi_1}|b|e^{j\phi_2}$, for which $\log F(s)$ has a constant real part if

$$\mathcal{R} \log F(s) = \mathcal{R}[\log ab + j(\phi_1 + \phi_2)] = \log a + \log b = \text{constant} \tag{E.3}$$

The condition that the sum of the logarithms of the lengths be constant is the same as that for constant potential of the electrostatic field; therefore, the solid lines in the plot are for constant magnitude of the function $\log F(s)$.

The lines of constant phase angle can be shown to cross the lines of constant magnitude at right angles by considering a small increment in $\log F(s)$.

$$d \log F(s) = \frac{d}{ds}[\log F(s)]\, ds \tag{E.4}$$

For a particular value of s, the derivative has a definite value. For $d \log F(s)$ to be real, ds must have some phase angle, say $\psi°$. For the increment $d \log F(s)$ to have a j component only, then the increment ds must be at right angles to the previous increment or have an angle $90° + \psi°$. For equal magnitudes of ds, the increments in $\log F(s)$ are likewise equal in magnitude; therefore, all the requirements for curvilinear squares are satisfied by the plot of $\log F(s)$ in the s plane.

The analogy between the complete root-locus plot and an electrostatic field is useful in sketching the root-locus plot. After the start of the locus is established near a pole, the remainder of the locus is the path which a charged particle would take in an electrostatic field, being repelled from the poles and attracted toward the zeros. The mid-point between points A and B in Fig. E.1 is unusual in that the field intensity of the field plot is zero at this point. The derivative of $\log F(s)$ is also zero; therefore, a large increment in s is needed to produce a given change in the function $\log F(s)$ in this region.

APPENDIX F

NOISE IN TERMS OF FREQUENCY COMPONENTS

Noise can be treated as the sum of specific frequency components in which only the mean-square amplitudes are significant. The transfer function of the servo must, in general, be selected as a compromise between transmission of noise and failure to transmit signal. The procedure described in this section requires that the optimum function be determined by a trial-and-error process.

Noise can be defined as unwanted signal. In a new experimental setup it will, of course, be recognized that many causes of noise voltages between signal terminals can be eliminated by reasonable care. Thus a ground-wire return, which is common to other circuits, should be avoided because the IR drops in the common wire appear as part of the signal. Similarly, shielding of the signal wire will reduce pickup due to capacitive coupling. Even after these precautions are taken, however, some other effect, such as vibration in a position pick-off, may continue to produce a noise signal which cannot be directly separated from the true signal. The reduction of this noise in the output of the system must then be based upon the difference in characteristics of the noise and the signal.

The description of noise in terms of frequency components is desirable because of the convenience of describing the transmission characteristics of the system on the same basis. A given sample of noise record can be described by a Fourier series:

$$f(t) = A_1 \sin \omega t + A_2 \sin 2\omega t + \cdots + B_1 \cos \omega t + \cdots \tag{F.1}$$

The amplitude of a particular component is determined by multiplying the above function of time by a sine wave of the frequency of that particular component and integrating over the interval T:

$$A_1 = \frac{2}{T} \int_0^T f(t) \sin \omega t \, dt \qquad T = \frac{2\pi}{\omega} \tag{F.2}$$

The above integral determines the fundamental component of the series in that one cycle of the sine wave is contained in the interval considered.

If an adjacent interval of the noise record, of equal length, were expressed as a Fourier series, the series would be different from the first one because the actual function of time is different. A common feature of all intervals of noise records, however, is that the mean-square values of the amplitudes are equal. The mean-square value of a Fourier series is the sum of the mean-square values of the individual amplitudes:

$$\tfrac{1}{2}[A_1^2 + A_2^2 + \cdots + B_1^2 + B_2^2 + \cdots] \tag{F.3}$$

This result can be verified by multiplying the series given in (F.1) by itself and noting that all the products involving different frequencies will have a mean value of zero. The factor $\frac{1}{2}$ arises in finding the average value of either a $\sin^2$ or $\cos^2$ term.

Consider now the description of the complete record of two adjacent intervals by a single Fourier series. The integral for a single term is split into two integrals; treating the A values as positive numbers, they may either add or subtract.

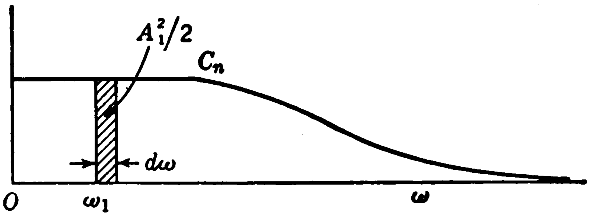


FIG. F.1. Noise in terms of frequency components.

$$\begin{aligned} A_1'' &= \frac{1}{T}\left[\int_0^T f(t)\,(\sin \omega t)\,dt + \int_T^{2T} f(t)\,(\sin \omega t)\,dt\right] \\ &= \tfrac{1}{2}(A_1 + A_1') \text{ or } \tfrac{1}{2}(A_1 - A_1') \end{aligned} \tag{F.4}$$

The mean value of the squares of the two possible results, however, is equal to the sum of the individual squares:

$$A_1''^2 = \tfrac{1}{2}\left[\frac{(A_1 + A_1')^2}{4} + \frac{(A_1 - A_1')^2}{4}\right] = \frac{A_1^2 + A_1'^2}{4} \tag{F.5}$$

The sum (F.3) of all the terms for the specific frequencies previously considered will be only half of that for the complete interval. But the double interval has a fundamental frequency which is half that of the original interval, and therefore the Fourier series will have twice as many terms. The mean-square amplitude of the noise record, as obtained by the sum of the squares of the amplitudes of all frequency components present, will therefore be the same for the double interval as for the original interval.

An extension of the concept of doubling the interval of noise record and thus doubling the number of frequency components present leads to the picture of noise distribution shown in Fig. F.1. The mean-square value of noise is the area under the curve, $\int_0^\infty C_n\,d\omega$. The abscissa is frequency plotted in a linear scale. The noise $A_1^2/2$ at a particular frequency ω_1 must specify a band width $d\omega$, such as 1 cps, to obtain a definite area. The ordinate to such a curve C_n has the dimensions of (noise amplitude)^2/cps.

If the transmission characteristic of the servo could be selected for a given frequency independently of the transmission at an adjacent frequency, then the choice of each value would be relatively simple. Thus let $C_n\,d\omega$ be the noise input at a particular frequency band and W be the transmission factor for that band; then $C_n\,d\omega\,W^2$ is the noise in the output. Similarly if $C_f\,d\omega$ is the signal content at that frequency band, then $C_f\,d\omega\,(1 - W)^2$ is the error due to failure to transmit the signal. Note that W should be a real number with no phase angle in order for the value of $1 - W$ to be a minimum for a given magnitude of W. The value of W for minimum error in the output is then determined from the condition

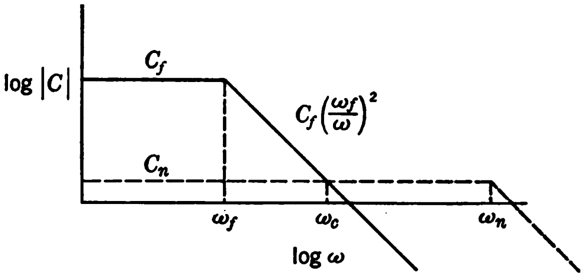


FIG. F.2. Frequency spectra of signal and noise.

$$\frac{d}{dW}\left[C_n\,d\omega\,W^2 + C_f\,d\omega\,(1 - W)^2\right] = 0 \tag{F.6}$$

Solving for W gives

$$W = \frac{C_f}{C_n + C_f} \tag{F.7}$$

Consider the special case of a system for which the frequency spectra of the signal C_f and noise C_n are as shown in Fig. F.2. The desired transfer function of the servo for minimum over-all error is, from (F.7), essentially unity up to the frequency ω_c for which the signal drops to the level of the noise. Above this frequency, the transfer function W should drop off as $(\omega_c/\omega)^2$. An approximate value of the over-all error in the output is simply the noise C_n times the frequency ω_c at which the signal and noise amplitudes are equal.

Unfortunately the transfer function at one frequency cannot be made independent of that at an adjacent frequency. Even the phase angle of the transfer function is correlated with the rate of change of amplitude with frequency. In the given problem, therefore, the choice of the function to obtain minimum total error is a difficult one. The form of the function is often determined from other considerations, however, and the choice of some parameters is the only problem. Moreover, the accuracy of the data regarding frequency spectrum of either noise or signal is often not great, and therefore high accuracy in determining optimum transfer functions is not justified. Consider again the case of the system whose inputs are shown in Fig. F.2. The transfer function is simply $1/(1 + Ts)$, in which T is selected to be $1/\omega_c$. The transfer function involved in failure to transmit the true signal is then

$$1 - W(s) = 1 - \frac{1}{1 + Ts} = \frac{Ts}{1 + Ts} \tag{F.8}$$

The corresponding frequency spectra of the error signal are then as shown in Fig. F.3. The plots are made with each axis logarithmic for

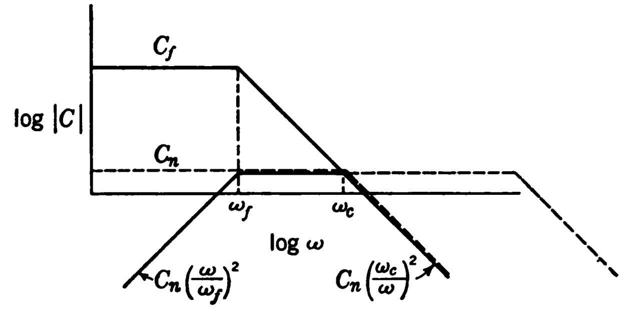


FIG. F.3. Transmission of signal and noise.

convenience in showing decreases of amplitude proportional to frequency squared. The area evaluation of total noise would require that the scales be linear on both axes. Calculations can be readily made, however, for simple cases like the one shown. Thus the noise content above the frequency ω_c is

$$\int_{\omega_c}^{\infty} C_n \frac{\omega_c{}^2}{\omega^2}\, d\omega = C_n\omega_c{}^2 \left[\frac{-1}{\omega}\right]_{\omega_c}^{\infty} = C_n\omega_c \tag{F.9}$$

The use of a transfer function $1/(1 + Ts)^2$, as indicated by the result in (F.7), would be a definite improvement, in that noise content above ω_c would then be approximately

$$\int_{\omega_c}^{\infty} C_n \left(\frac{\omega_c}{\omega}\right)^4 d\omega = C_n\omega_c{}^4 \left[\frac{1}{-3\omega^3}\right]_{\omega_c}^{\infty} = \frac{1}{3}\, C_n\omega_c \tag{F.10}$$

the error due to failure to transmit the signal, however, would be increased. The determination of the optimum transfer function would be a trial-and-error process as here presented.

APPENDIX G

AUTOCORRELATION FUNCTIONS

The response of a system can be expressed in terms of the input signal and the weighting function of the system by a convolution integral. A product of such integrals gives the mean-square value of the response, which depends upon the autocorrelation function of the signal.

The problem of servo design considered in the main portion of this book is to make the error of the system as small as possible for the expected input signals, consistent with the general limitations of stability and complexity. In general, however, the input signal is mixed with a noise signal, and therefore the servo must also act to filter out the noise. The initial developments in this appendix apply to any such problem but will later be applied specifically to the problem of selecting a parameter of a servo function which will have the minimum over-all error from transmission of noise and failure to transmit signal.

As explained in Art. 11.8, the output of a system whose response to a unit impulse is $w(t)$ can be expressed by a convolution integral:

$$x(t) = \int_0^{t_1} f(t_1 - t)w(t)\,dt \qquad \text{(G.1)}$$

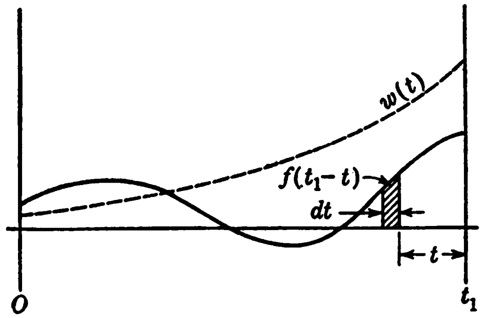


Fig. G.1. Convolution integral.

This integral expresses the total response as the sum of the transients due to each impulse of the input, as shown in Fig. G.1. The time t is the age of the impulse, or the time between the occurrence of the impulse and the present time t_1. The square of the output of a system is the product of two such convolution integrals.

$$x^2(t_1) = \int_0^{t} f(t_1 - u)w(u)\,du \int_0^{t_1} f(t_1 - v)w(v)\,dv \qquad \text{(G.2)}$$

The independent variables u and v are used to represent age of the signal in the two integrals for convenience in later manipulations. The time function $f(t_1 - t)$ is not known; the object is to determine the probable value of $x^2(t_1)$ in terms of some characteristic of $f(t_1 - t)$.

One way to picture this product of integrals is shown in Fig. G.2. Time zero is shown in the upper left-hand corner. The weighting function $w(u)$ is plotted to the right; the weighting function $w(v)$ is plotted down. Along any one column, as at time u, the impulse of the signal times the ordinate of the weighting function is the part of the response due to that impulse. Similarly, for any one row, the product of the input impulse and weighting-function ordinate is part of the response. At the intersection of each row and column is a product of four numbers,

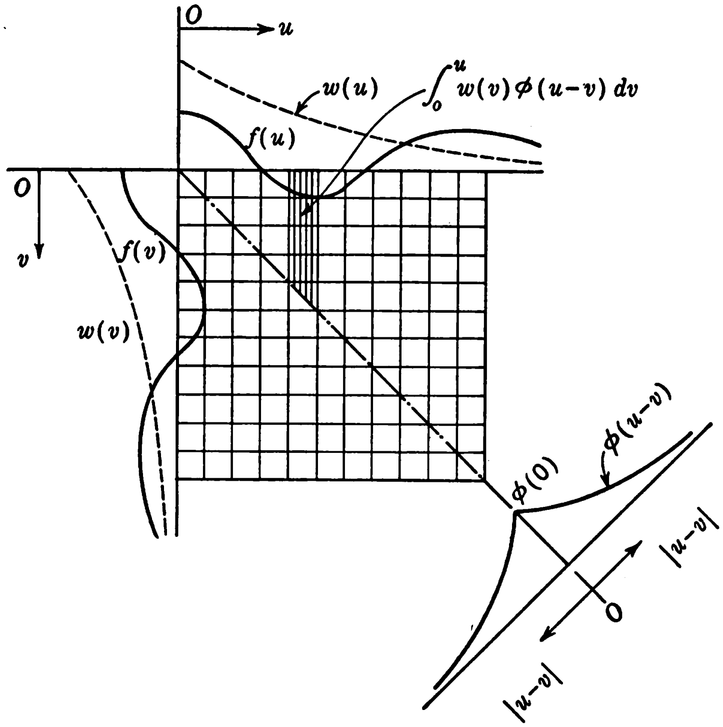


FIG. G.2. Elements of mean error squared.

which is one element of the total response squared. The total response squared is the sum of the elements in the entire square array. The time increment du or dv can be associated with the weighting function in each case rather than with the input signal. At any element along the principal diagonal from upper left to lower right, the input signal, at any time, is multiplied by itself. The probable value of this product is, therefore, simply the root-mean-square value of the signal, denoted by $\phi(0)$. Along any diagonal parallel to the principal diagonal, an input will always be multiplied by an input which is offset by fixed time interval. The probable value of this product will always be less than that for which no offset is present and is a function of the offset time. The curve plotted along the section across these diagonals has the offset time identified as $u - v$ because u is the age of one input signal and v is the age of the other input signal. The curve is symmetrical about zero offset time. This curve is called the autocorrelation function of the signal, because it describes the correlation of the input with itself as a function of offset time. One way of finding the total probable error is to sum, or integrate, along columns and then along rows. The integral for the shaded column in Fig. G.2 is

$$\int_0^u \phi(u - v)w(v)\,dv \tag{G.3}$$

The variable u is constant along this column, and the integration is carried out from $v = 0$ at the top to $v = u$ at the diagonal. Each of the column integrals should then be multiplied by the corresponding weighting impulse $w(u)\,du$ at the time u and the integration carried out from $u = 0$ at the left to $u = t$ at the right:

$$\overline{x^2(t)} = 2 \int_0^t w(u)\,du \int_0^u \phi(u - v)\,w(v)\,dv \tag{G.4}$$

For a specific value $t = t_1$ the probable value of the error squared for that length of input signal is calculated. If t is allowed to approach infinity, then the probable value $x^2(\infty)$ is determined; this is a finite value if the system is one with damping present for all modes.

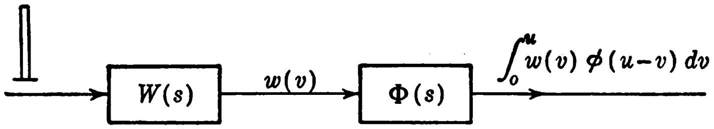


FIG. G.3. Partial block diagram for generation of mean error squared.

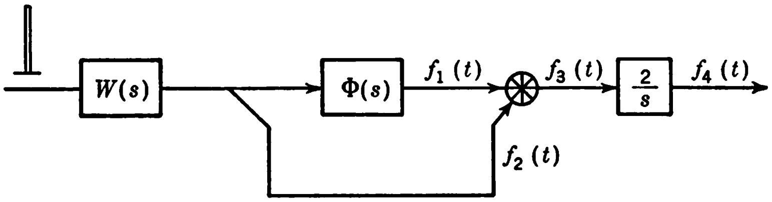


FIG. G.4. Block diagram for mean error squared.

The factor 2 appears in (G.4) because the integration below the diagonal will yield the same result as the integration above the diagonal because of the symmetry throughout. Another picture of the integral given in (G.4) is obtained by recognizing that the integral in (G.3) represents a standard convolution integral for a system whose transfer function is $\Phi(s)W(s)$ in response to a unit impulse, as shown by the block diagram in Fig. G.3. The unit impulse acting on $W(s)$ produces the weighting function $w(v)$, which in turn produces the output given in Eq. (G.3) after transmission through $\Phi(s)$. The double integral in (G.4) involves multiplying the output of the block diagram in Fig. G.3 by $w(u)$, as indicated by the symbol $\circledast$ in Fig. G.4.

The evaluation of the output of this block diagram can be made in any of several ways, depending upon the problem. Consider first the case in which the system is an undamped resonant one:

$$W(s) = \frac{1}{1 + s^2/\omega_1^2} \qquad \text{or} \qquad w(t) = \omega_1 \sin \omega_1 t \tag{G.5}$$

The input is a noise signal whose root-mean-square amplitude of the noise is 1, but the fluctuations are so rapid that the value at one instant is not related to the value at an adjacent instant. The autocorrelation function $\Phi(u - v)$ is then a unit impulse, and the transfer function $\Phi(s)$ is simply 1. The function $f_1(t)$ is therefore equal to $f_2(t)$, which is $\omega_1 \sin \omega_1 t$. The expression for $f_4(t)$ is therefore

$$f_4(t) = \int_0^t 2\omega_1^2 \sin^2 \omega_1 t \, dt = \omega_1^2 t - \tfrac{1}{2}\omega_1 \sin 2\omega_1 t \tag{G.6}$$

A graph of this function in Fig. G.5 shows the pattern in which the square of the output, which is error in this case, will probably build up with time. Initially the rate of change of the error squared is zero, but after $\omega_1 t = 2\pi$, the error squared builds up approximately linearly with time. This transient solution for the error-squared function is one of the advantages of the autocorrelation-function type of approach.

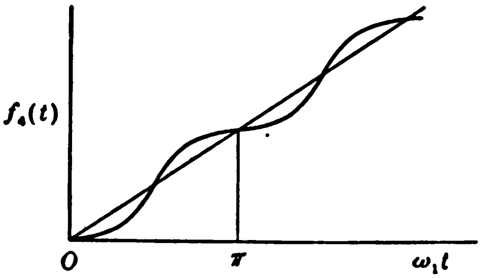


FIG. G.5. Transient build-up of mean error squared for resonant system with $\Phi(s) = 1$.

Consider now the problem of selecting a parameter of a transfer function to have minimum over-all error squared due to transmission of noise and failure to transmit signal. Assume that the transforms of the autocorrelation functions of the noise and signal are

$$\Phi_n(s) = \frac{\phi_n(0)}{s - s_n} \qquad \Phi_f(s) = \frac{\phi_f(0)}{s - s_f} \tag{G.7}$$

The transfer function from input to output will be selected to be simply

$$W(s) = \frac{1}{1 + Ts} = \frac{1/T}{1/T + s} = \frac{-s_c}{s - s_c} \tag{G.8}$$

The problem is simplified to choosing merely the value of the time constant T or the pole s_c. Any noise present in the output will be part of the error and can be evaluated from the block diagram in Fig. G.6.

The function $f_1(t)$ of time is

$$f_1(t) = \frac{-s_c \phi_n(0)}{s_c - s_n} (e^{s_c t} - e^{s_n t}) \tag{G.9}$$

Multiplication by $f_2(t) = -s_c e^{s_c t}$ gives $f_3(t)$:

$$f_3(t) = \frac{s_c{}^2 \phi_n(0)}{s_c - s_n} \left(e^{2s_c t} - e^{(s_n + s_c)t}\right) \tag{G.10}$$

Integration determines $f_4(t)$:

$$f_4(t) = \frac{2s_c{}^2 \phi_n(0)}{s_c - s_n} \left[\frac{e^{2s_c t}}{2s_c} - \frac{e^{(s_c + s_n)t}}{s_n + s_c}\right]_0^t \tag{G.11}$$

Evaluation of the limiting value of $f_4(t)$ for the case in which s_c and s_n are both negative real numbers gives

$$f_4(\infty) = \frac{-2s_c{}^2 \phi_n(0)}{s_c - s_n} \left[\frac{1}{2s_c} - \frac{1}{s_n + s_c}\right] = \frac{\phi_n(0) s_c}{s_c + s_n} \tag{G.12}$$

Any problem could be carried through by this process of multiplication, and integration, but some theorems from Laplace transform theory permit simplification of the calculations. The first of these is the complex convolution theorem:

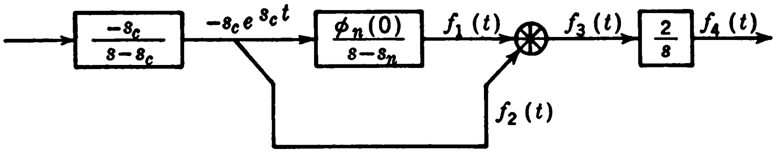


FIG. G.6. Block diagram for system with simple filter.

$$F_3(s) = \mathcal{L} f_1(t) f_2(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F_1(s) F_2(s' - s)\, ds \tag{G.13}$$

Note that $f_3(t)$ in Fig. G.6, for example, is expressed as the instantaneous product of two time functions. Therefore, Eq. (G.13) is a means of obtaining the transform of this function, or $F_3(s)$. The functions $F_1(s)$ and $F_2(s)$ are the transfer functions for which a unit-impulse input will produce $f_1(t)$ and $f_2(t)$, respectively. Each side of (G.13) can be easily calculated for the simple case of

$$f_1(t) = A_1 e^{s_1 t} \qquad f_2(t) = A_2 e^{s_2 t} \tag{G.14}$$

The direct evaluation of the transform of the product gives

$$F_3(s) = \mathcal{L} A_1 e^{s_1 t} A_2 e^{s_2 t} = \mathcal{L} A_1 A_2 e^{(s_1 + s_2)t} = \frac{A_1 A_2}{s - (s_1 + s_2)} \tag{G.15}$$

The integral in the complex plane, given in the right side of (G.13), is to be evaluated for the case of

$$F_1(s) = \frac{A_1}{s - s_1} \qquad \text{and} \qquad F_2(s) = \frac{A_2}{s - s_2}$$

$$F_3(s) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} \frac{A_1}{s - s_1} \frac{A_2}{s' - s - s_2}\, ds \tag{G.16}$$

The poles of the functions are shown in Fig. G.7; the path of integration specified in this theorem must be such as to separate the poles of the $F_2(s' - s)$ function from those of the $F_1(s)$ function. The j axis is usually acceptable ($c = 0$), because the poles of the original functions are all in the left half plane, making the poles of the $F_2(s' - s)$ function lie in the right half plane. The integral can be completed around the right half plane at infinite radius without changing its value if the value of the function is zero along this path. The integral can then be calculated simply as the residue at the enclosed pole. The sign is changed because the encirclement is clockwise rather than counterclockwise as for the left half plane.

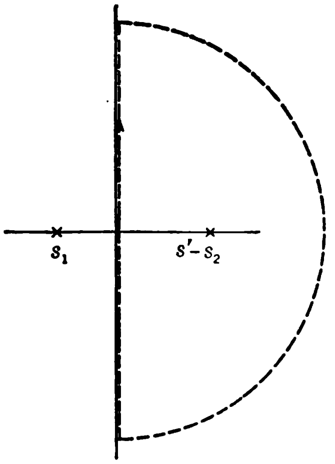


FIG. G.7. Integration in the complex plane.

$$F_3(s) = \left[\frac{A_1}{s - s_1}\right]_{s=s'-s_2} A_2 = \frac{A_1 A_2}{s' - s_1 - s_2} \tag{G.17}$$

Note that this result checks the value obtained in (G.15). The variable s' in (G.17) has the same meaning as s in (G.15), because s' was introduced in (G.13) only for convenience in expressing the complex convolution integral.

In the general case, $f_1(t)$ and $f_2(t)$ are both a series of terms of the form Ae^{st}. The evaluation of $F_3(s)$ could then be made by either of the methods just described in exactly the same manner. The result in each case would be a series of terms of the form of (G.15) involving every combination of the terms in the $F_1(s)$ series with those of the $F_2(s)$ series.

The final part of the evaluation of $f_4(t)$ is to let time go to infinity. This result can be obtained in most cases simply by letting s go to zero in $F_3(s)$, as can be shown by the following reasoning: The transform $F_4(s)$ is $(2/s)F_3(s)$. The constant term in $f_4(t)$ is found by omitting the s factor from the denominator of $F_4(s)$ and letting $s = 0$ in the rest of the function. For a stable system all terms in $f_4(t)$ other than the constant term go to zero for an infinite value of time. The final value of $f_4(t)$ is found simply by letting $s = 0$ in $2F_3(s)$. An exception is for $F_3(s)$ to have the factor s.

The evaluation of the error resulting from failure to transmit the signal properly can now be readily calculated. The transfer function $W(s)$ of the system from input to output has been specified in (G.8) to be $-s_c/(s - s_c)$. The proper transfer function for no error between input and output would be 1. The evaluation technique just developed requires, however, that 1 be approximated as $-s_d/(s - s_d)$ in order that the function may have a zero value at $s = \infty$, as required by the contour integration. After $F_3(s)$ is obtained, s_d can be allowed to approach infinity. The block diagram for error due to failure to transmit signal properly then is given in Fig. G.8.

Evaluating the $F_1(s)$ function $[-s_d/(s - s_d) + s_c/(s - s_c)]\phi_f(0)/(s - s_f)$ at the poles of the $F_2(s' - s)$ function,

$$F_3(s') = [F_1(s)]_{s=s'-s_d}(-s_d) - [F_1(s)]_{s=s'-s_c}(-s_c) \tag{G.18}$$

Let s' be zero in $F_3(s)$ to find $f_4(\infty)$.

$$f_4(\infty) = 2\left\{\left(\frac{-s_d}{-2s_d} + \frac{s_c}{-s_d - s_c}\right)\left[\frac{\phi_f(0)}{-s_d - s_f}\right](-s_d) - \left(\frac{-s_d}{-s_c - s_d} - \frac{-s_c}{-2s_c}\right)\left[\frac{\phi_f(0)}{-s_c - s_f}\right](-s_c)\right\} \tag{G.19}$$

Let s_d approach infinity, which permits simplification to

$$f_4(\infty) = 2\left\{\left(\frac{1}{2} + 0\right)[\phi_f(0)] - \left(1 - \frac{1}{2}\right)\left[\frac{\phi_f(0)s_c}{s_c + s_f}\right]\right\} = \frac{\phi_f(0)s_f}{s_c + s_f} \tag{G.20}$$

Note the result in (G.20) is of the same form as that in (G.12) for the noise transmission except that the numerator is s_f rather than s_c. This difference is very significant, however, as can be seen from an approximate plot of the total error squared as a function of the choice of s_c, as shown in Fig. G.9. Both axes are plotted with logarithmic scales to permit straight-line approximations for the value of the functions versus s_c. The minimum value of the sum of the two sources of error squared is at the point of intersection of the two lines. The optimum value of $s_c = s_m$ can thus be determined from

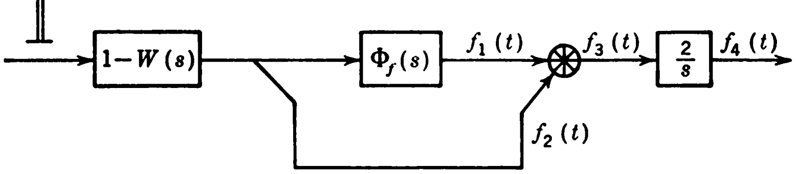


FIG. G.8. Block diagram for error in transmission of signal.

$$\phi_f(0)\frac{s_f}{s_m} = \phi_n(0)\frac{s_m}{s_n}$$

or

$$s_m = \sqrt{\frac{\phi_f(0)}{\phi_n(0)}s_f s_n} \tag{G.21}$$

Solving for the minimum value of the sum gives twice the value of either term:

$$\phi_m = 2\frac{\phi_n(0)s_m}{s_n} = 2\sqrt{\phi_f(0)\phi_n(0)\frac{s_f}{s_n}} \tag{G.22}$$

Note that the plot shown in Fig. G.9 is a special case with the result given in (G.21) applicable only for this case. The intersection would occur for s_m greater than s_n if $\phi_n(0)/\phi_f(0)$ is less than s_f/s_n.

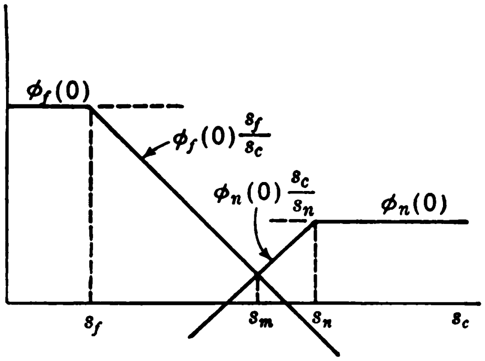


FIG. G.9. Choice of parameter s_c for minimum over-all error squared.

A comparison of the result of this analysis with those obtained from the frequency-response analysis of the same problem in Appendix F shows that the results are essentially the same. Thus the frequency band of the noise was assumed to be high in Appendix F, corresponding to s_n being very high. The noise level C_n in Fig. F.2 is proportional to $\phi_n(0)/s_n$ in autocorrelation-function terminology. Similarly the signal level in C_f in Fig. F.2 is proportional to $\phi_f(0)/s_f$. The signal drops off to noise level $C_n \cong \phi_n/s_n$ at the frequency ω_c

$$\left(\frac{\phi_f(0)}{s_f}\right)\left(\frac{s_f}{\omega_c}\right)^2 = \frac{\phi_n(0)}{s_n} \qquad \text{or} \qquad \omega_c = \sqrt{\frac{\phi_f(0)}{\phi_n(0)}\, s_n s_f} \tag{G.23}$$

Note that this value of ω_c corresponds to the value of s_m determined in (G.21).

The autocorrelation-function method of analysis can be correlated with the frequency-response method in more general terms by proper interpretation of the convolution integral (G.13). Thus if the integration is carried out along the j axis and if the value of s' is set equal to 0 (in G.13), the steady-state value of error squared due to transmission of noise is

$$\overline{\epsilon^2(\infty)} = \frac{2}{2\pi j}\int_{-\infty}^{+\infty} \Phi_n(-j\omega)W(j\omega)W(-j\omega)\,dj\omega \tag{G.24}$$

The function $W(j\omega)$ is the conjugate of $W(-j\omega)$, and therefore

$$W(j\omega)W(-j\omega) = |W(j\omega)|^2$$

$$\overline{\epsilon^2(\infty)} = \frac{2}{2\pi j}\int_{-\infty}^{+\infty} |W(j\omega)|^2\Phi_n(-j\omega)\,dj\omega \tag{G.25}$$

The evaluation of the integral for $\Phi_n(-j\omega)$ is now seen to give the same value as for $\Phi_n(+j\omega)$, and therefore the integral can be expressed as

$$\overline{\epsilon^2(\infty)} = \frac{1}{2\pi j}\int_{-\infty}^{+\infty} |W(j\omega)|^2[\Phi_n(j\omega) + \Phi_n(-j\omega)]\,dj\omega \tag{G.26}$$

The quantity in brackets is now a real number. The integral now has the form used in Appendix F, in which the noise at each frequency is transmitted according to the square of the absolute magnitude of the transfer function at that frequency. The integral from $-j\infty$ to 0 is the same as that from 0 to $j\infty$, and therefore (G.26) simplifies to:

$$\overline{\epsilon^2(\infty)} = \frac{2}{\pi}\int_0^\infty |W(j\omega)|^2 \mathcal{R}\Phi_n(j\omega)\, d\omega \tag{G.27}$$

A different description of the noise function can be obtained by keeping the convolution integral (G.13) in more general form than (G.24). Although s' is again zero for the steady-state evaluation, the variable s can be complex as long as the path of integration separates the poles of $F_1(s)$ from those of $F_2(-s)$.

$$\overline{\epsilon^2(\infty)} = \frac{2}{2\pi j}\int_{c-j\infty}^{c+j\infty} W(s)\Phi_n(s)W(-s)\, ds \tag{G.28}$$

$F_1(s)$ and $F_2(s)$ can be interchanged, which makes the integral the same as (G.28) except that $\Phi_n(s)$ is replaced by $\Phi_n(-s)$. The factor $W(s)W(-s)$ is present in both integrands, and therefore the integral can be expressed as

$$\overline{\epsilon^2(\infty)} = \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} W(s)W(-s)[\Phi_n(s) + \Phi_n(-s)]\, ds \tag{G.29}$$

The sum in brackets can be replaced by a product $\Psi_n(s)\Psi_n(-s)$ in which symmetry with respect to the sign of s is preserved.

$$\overline{\epsilon^2(\infty)} = \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} W(s)W(-s)[\Psi_n(s)\Psi_n(-s)]\, ds \tag{G.30}$$

The poles of $\Psi_n(s)$ must be the same as those of $\Phi_n(s)$. The zeros of $\Psi_n(s)$ can be found by a root-locus plot which makes the ratio

$$\frac{\Phi_n(s)}{\Phi_n(-s)} = -1$$

The plot has some unique features of symmetry as will be illustrated for the case

$$\frac{\Phi_n(s)}{\Phi_n(-s)} = \frac{(s-q_1)}{(s-p_1)(s-p_2)}\,\frac{(-s-p_1)(-s-p_2)}{(-s-q_1)} = -1 \tag{G.31}$$

The coefficient of s must be $+1$ for each factor to be treated as a vector to the s point in the plot. This modification results in a net change in sign on the right side of (G.31) and therefore the locus is for a net angle of 0° in Fig. G.10.

The amplitude factor is always 1 for this type of plot, a condition which is always satisfied for points along the j axis or at infinity. The only measurement needed on Fig. G.10 is the determination of the frequency

ω_1 at which the net angle of the vectors is zero. In general, roots of an equation of the form of (G.31) can be complex numbers, but the values need be determined only in one quadrant. The usual symmetry of the functions about the real axis makes the roots oocur as conjugates; the special symmetry about the j axis makes the roots symmetrical about this axis.

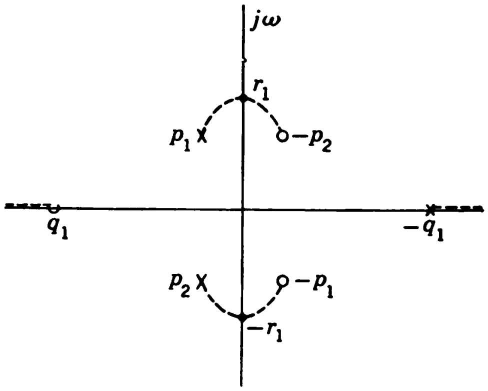


FIG. G.10. Root locus of $\Phi_n(s)/\Phi_n(-s) = -1$.

Knowing $\Psi(s)$ in factored form in (G.30) can be helpful in selecting the transfer function $W(s)$ which will produce a minimum error squared for a given noise function and a given signal to be transmitted. In evaluating (G.30) along the j axis, the phase angles are not required as they are in (G.27) to find the real part of $\Phi_n(s)$. Note, however, that the only zero of $\Psi_n(s)$ determined in Fig. G.10 is $j\omega_1$ and therefore $\Psi(-j\omega_1)$ is not the conjugate of $\Psi_n(j\omega)$ even though $\Psi_n(s)\Psi_n(-s)$ is a real number.

PROBLEMS

CHAPTER 1

1.1. Sketch a control system to maintain the water level in a reservoir at a fixed level for an arbitrary output rate but a controllable input rate.

1.2. Sketch a system for maintaining the total flow rate and temperature of a liquid with hot and cold sources of the liquid available.

1.3. Given the torsional pendulum shown in Fig. 3.1*b*, what is the error due to a velocity $\dot{\theta}_i$. What is the additional error for an acceleration $\ddot{\theta}_i$?

1.4. Given a system in which the input and output are related by the equation

$$\left(1 + B\frac{d}{dt}\right)\theta_i = \left(1 + T_1\frac{d}{dt}\right)\left(1 + T_2\frac{d}{dt}\right)\theta_o$$

What is the error required to maintain a constant output velocity $\dot{\theta}_o$? What is the additional error to cause the output to accelerate at the rate $\ddot{\theta}_o$?

★1.5.[1] For the system given in Prob. 1.4, what is the error required to maintain a sinusoidal oscillation sin ωt?

CHAPTER 2

2.1. Given the synchro system described in Art. 2.1, why does reversing the connection of two wires on either stator reverse the direction of rotation of the output with respect to the input?

2.2. For the pick-off shown in Fig. 2.4 sketch the approximate shape of the curve of output voltage versus angle of the coil when the coil makes a full pass through the field of the C cores.

2.3. Sketch the output of a 400-cycle demodulator due to a 60-cycle input.

2.4. Some small d-c servo motors have a series field rather than a shunt field so that the flux is proportional to the armature current. Sketch the approximate shape of the speed-torque curves for full voltage and one-half voltage.

★2.5. An induction motor has an approximate input impedance of the form

$$\left(\frac{R}{s}\right) + jX$$

in which X is the reactance, numerically equal to R, and s is the slip, which varies from 1 for stall to 0 for top speed. If the output resistance of the amplifier is also equal to R, the voltage impressed on the motor is appreciably affected by motor speed. Sketch the approximate speed-torque curve and compare with the curve given in Fig. 2.14.

2.6. Sketch the approximate curves of force versus gap for the solenoid in Fig. 2.16 when an a-c voltage rather than a d-c voltage is impressed across the coil.

★2.7. Making all the approximations necessary to achieve a simple result, draw the circuit diagram which is equivalent to the hydraulic system shown in Fig. 2.17. Assume that the load on the actuator is an inertia only.

---

[1] A star is used to identify problems which are more difficult, in that some extension of the theory presented in the text is required for their solution.

***2.8.** Include the effects of coulomb friction in (2.13) and solve for the optimum gear ratio. What is the result for the limiting case of m_F being very large?

2.9. Assume that the tangential force per square inch of rotor surface remains constant when the dimensions of a motor are reduced. If a small motor is made with each dimension one-half that of a large motor, what is the torque of the small motor compared to that of the large motor? What is the maximum acceleration of the small motor compared to that of the large motor?

2.10. The a-c induction-motor characteristics of Fig. 2.14 are shown in Fig. P2.1 with typical values given. The motor rotor inertia $J = 3$ g-cm^2. Find the following:

(*a*) Maximum acceleration in radians per second per second.

(*b*) Damping constant $D = \Delta m/\Delta n$ for line B; for line C.

(*c*) "Time constant" of the motor, here defined as J/D for line B; for line C.

(*d*) What is the approximate value of the maximum power output of this motor in watts? At approximately what speed is this power developed?

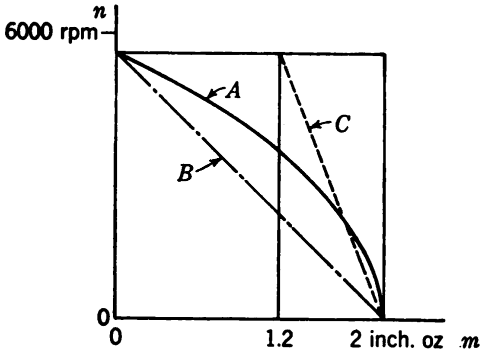


FIG. P2.1. Torque-speed characteristics of an a-c induction motor with linear approximations.

***2.11.** A motor with a 2-in.-oz maximum torque and a rotor inertia of 3 g-cm^2 is to be used for the purpose of accelerating a load of 30,000 g-cm^2 as rapidly as possible.

(*a*) Neglecting static friction and assuming perfect gears, what gear ratio should be used, and what is the acceleration of the load which is possible?

(*b*) If double the optimum gear ratio were used, what would be the maximum acceleration?

(*c*) If static friction at the load were 50 in.-oz and the gear efficiency 75 per cent, and if the gear ratio selected in part *a* were used, what would be the maximum acceleration?

(*d*) For the conditions given in part *c*, except for the gear ratio, what gear ratio should be used, as judged from the curve in Fig. 2.20, and what would be the maximum acceleration?

CHAPTER 3

3.1. A block diagram can be drawn for any system by following the sequence of cause and effect. Demonstrate by making a block diagram of the sort in Fig. 3.1*a* for the spring-mass system shown in Fig. 3.1*b*.

3.2. Draw a figure similar to Fig. 3.3 for an electrical system and for a thermal system, each having a single form of energy storage. In each case identify the time constant.

***3.3.** The value of e is often specified in calculus books as lim $(1 + x)^{1/x}$ as x approaches 0. Show that this definition agrees with the definition that e is the value such that e^y has a slope numerically equal to its value at any point.

3.4. How many time constants are needed for an exponential to decay to the following fractions of its original value: $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{10}$, and $\frac{1}{100}$?

3.5. What is the value of $\sqrt{j}$?

3.6. The advantage of using the polar form of complex numbers for multiplication can be demonstrated by solving a particular example each way. Evaluate the product $(1 + j0.1)(1 + j0.5)(1 + j1)(1 + j2)$ by conversion of each factor to polar form and by direct complex algebra. The conversion to polar form may be made by direct plotting and measurement if desired.

3.7. Find the output of the system in Fig. 3.14 for the initial conditions that $\theta_o = 0$ and $d\theta_o/dt = v_o$.

***3.8.** Draw the vectors to represent the force involved in (3.47) and show that the value of s given in (3.51) is correct because the vector sum of the forces is zero.

CHAPTER 4

4.1. The block diagram for the system in Fig. 4.1 can be developed in complete detail by showing individual loading effects as separate feedback signals, as shown in Fig. P4.1 for the generator field. Complete the diagram in Fig. 4.3 for the system in Fig. 4.1 including identification of motor torque as a separate variable. Motor inertia is J_m and load inertia is J_L.

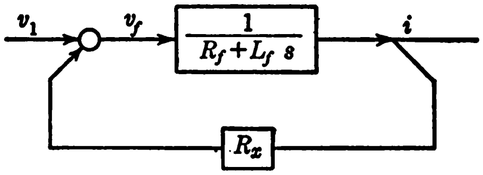


FIG. P4.1. Block diagram including the effect of output impedance of an amplifier.

4.2. The transfer function from input θ_i to output θ_o is $1/(1 + Ts)$. Find the input needed to produce the output $e^{\sigma t} \sin (\omega t)$. What is the condition for the phase shift from input to output to be 90°?

***4.3.** The transfer function from input θ_i to output θ_o is $1/(1 + Ts)$. The input is an impulse of unit area starting at $t = 0$ but with finite width Δt. Solve for the exact expression for the output after the impulse. The effect of a finite width is to change the amplitude of the transient by a fixed factor compared to the transient for an instantaneous impulse. What is the numerical value of this factor for $\Delta t = T$?

***4.4.** Devise another argument to show that the output for a given input is independent of the order of the blocks in series.

4.5. Write the transient output for a unit-impulse input acting on a system for which the transfer function is

$$\frac{\theta_o}{\theta_i}(s) = \frac{1 - s/q_1}{(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)}$$

***4.6.** What is the transfer function needed to generate a unit sine wave from a unit impulse?

4.7. Modify Fig. 7.11 to include initial conditions of $d\theta_o/dt = v(0)$ and $\theta_f = \theta_f(0)$.

4.8. Solve for the output of the system in Fig. 4.9 due to the initial conditions shown.

4.9. Set up a block diagram showing the relationship between v, i, m, and n using Eqs. (2.1) and (2.5) for a load having an inertia J.

4.10. Set up a block diagram showing the signals v, i, dx/dt, and f using Eqs. (2.8) and (2.9).

CHAPTER 5

5.1. Apply Routh's criterion to Eq. (5.7) and determine the limiting condition for the system to be stable. What is the natural frequency of the roots?

5.2. Sketch the phase-shift and attenuation characteristics of $1/(1 + Ts)$ for $s = j\omega$ from $\omega = 0$ to $\omega = 5/T$.

***5.3.** The inverse transfer function for a quadratic system in dimensionless form is $1 + 2\zeta s/\omega_0 + s^2/\omega_0^2$, in which ζ is the damping ratio and ω_0 is the undamped natural frequency.

(*a*) Solve for the ratio ω_n/ω_0 as a function of ζ, where ω_n is the natural frequency of the roots.

(*b*) Solve for the ratio ω_m/ω_0 as a function of ζ, where ω_m is the frequency at which the magnitude of the direct transfer function is a maximum.

★5.4. Show that the vector plot of the frequency response of the function

$$\frac{1 + \alpha Ts}{1 + Ts}$$

is a circle with a diameter between α and 1 on the real axis. The result can be achieved simply either by dividing numerator by denominator or by establishing first the locus of $1/(1 + Ts)$ and then constructing the given function.

5.5. (*a*) Show the point on the inverse vector plot and the direct vector plot for the closed-loop response to be 1.4 in magnitude and 60° in phase shift.

(*b*) If this point is to be the maximum magnification, show the tangent to the open-loop curve in each plot.

5.6. For active compensation, the desired inverse function of a loop is simply equal to the direct transfer function of the feedback path. Figure 5.15 shows the actual, as well as the desired, inverse function. Sketch the ratio of $(\Theta_i/\Theta_o)(j\omega)$ to $(\Theta_f/\Theta_o)(j\omega)$.

5.7. The conversion from numbers to logarithms to the base 10 has many convenient approximations based on the one fact that $\log_{10} 2 = 0.3$. (*a*) Find the values of $\log_{10}$ of $2^{\frac{1}{2}}$, 2, 2^2, $\frac{10}{2}$ and $10/2^{\frac{1}{2}}$. (*b*) Plot the curve of $\log_{10} a$ versus a for values of a from 1 to 10, using the points determined in (*a*).

5.8. Using approximations of the type used in Fig. 5.18, sketch the transfer function $(\Theta_i/\Theta_e)(j\omega)$ from the plot of $(\Theta_o/\Theta_e)(j\omega)$ shown in Fig. 5.18.

★5.9. An optional method of obtaining the closed-loop function from the open-loop function makes use of the polar plot with logarithmically spaced circles shown in Fig. P5.1. The inverse open-loop function is plotted from the logarithmic diagram on a transparent sheet. Shifting the sheet to the right by an amount proportional to the gain K then gives the values of the inverse closed-loop plot, which can be read in logarithmic terms versus frequency. Use this method to obtain a more accurate plot of the closed-loop function in Fig. 5.18.

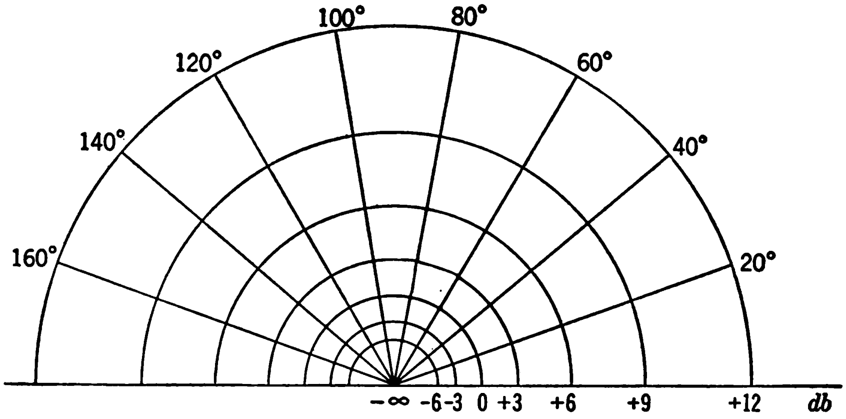


FIG. P5.1. Logarithmic-polar conversion chart.

5.10. A lead network is frequently needed to obtain a given phase lead at a particular frequency, but the corresponding amplification of a high-frequency signal compared to that for a d-c signal should be minimized.

(*a*) If a single lead network is used, what is the amplification needed to obtain 1 rad phase lead?

(*b*) What amplification is needed if an infinite number of lead networks in series are allowed in obtaining 1 rad phase lead?

5.11. Given a system like that shown in Fig. 5.1 with $T_1 = 1$ sec and $T_2 = \frac{1}{4}$ sec.

(*a*) What is the value of the gain factor K which would cause the system to maintain a continuous oscillation?

(*b*) What is the value of K which will make the maximum value of $(\Theta_o/\Theta_i)(j\omega)$ equal to 1.4 for any frequency?

★5.12. Given the system in Prob. 5.11, the frequency for a maximum $(\Theta_o/\Theta_i)(j\omega)$ is to be raised.

(*a*) What can be achieved by use of a single lead network having an α factor of $\frac{1}{10}$?

(*b*) If the system in Prob. 5.11 could take a sudden change in θ_i of 0.1 rad without saturation of the output stage of the amplifier, what step change in θ_i will produce saturation after the lead network is added and the gain adjusted?

CHAPTER 6

6.1. The inverse vector plots for the quadratic function $1 + 2\zeta s/\omega_0 + s^2/\omega_0^2$ are readily calculated and plotted as a complex number. Sketch these curves for $\zeta = 0.1$, 0.3, 0.5, 0.7, and 1 with values of ω up to $2\omega_0$.

6.2. Sketch the frequency response of the system whose open-loop transfer function is $K(1 - Ts)/(1 + Ts)s$.

(*a*) Use both a direct and inverse plot, and use gain values which make the curves cross the negative axis both inside and outside the -1 point.

(*b*) Identify which curve corresponds to a stable system in each case, and justify the conclusion on the basis of net rotations of a vector pivoted at the -1 point.

★6.3. Sketch the frequency response of the system whose open-loop transfer function is $-K(1 + Ts)/(1 - Ts)s$.

(*a*) Use both direct and inverse plots with values of gain which cause the curves to cross the negative real axis both inside and outside the -1 point.

(*b*) Identify which curve corresponds to a stable system, and justify the conclusion on the basis of net rotations of a vector pivoted at the -1 point.

6.4. Sketch the conformal map of the function $(1 - s/r_1)(1 - s/r_2)$ for the grid of lines in the s plane shown in Fig. P6.1.

6.5. (*a*) Plot the vector curve for

$$\Theta_e/\Theta_o(j\omega) = (1 + s)(1 + s/3)s/K$$

for $K = 2$. (*b*) Find the dominant roots by sketching a conformal map. (*c*) Cancel out the transfer function corresponding to the assumed quadratic and determine the remaining part of the transfer function.

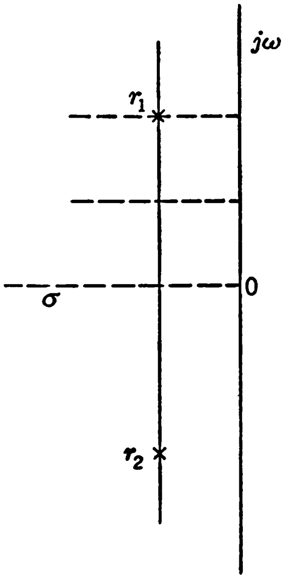


Fig. P6.1. Square grid in the s plane.

6.6. Given the conformal map shown in Fig. 6.8, find the amplitude of the transient terms represented by these roots for a unit-step input.

★6.7. By successive approximations obtain the approximate values of the roots of the system shown in Fig. 6.12.

★6.8. A square grid of lines in the s plane will map into a curvilinear square grid in the $F(s)$ plane. Justify or prove that a square grid of lines in the $F(s)$ plane will map into a curvilinear square grid in the s plane.

CHAPTER 7

7.1. The zeros and poles of several transfer functions are shown along the real axis in Fig. P7.1.

(*a*) Denote the portions of the locus of roots along the real axis by dotted lines just above the axis.

(*b*) Estimate the positions at which the loci break away from the real axis.

***7.2.** A root-locus plot frequently has points widely separated, as for the equation $K(1 + s/2)/s^2(1 + s/40)(1 + s/100)(1 + s/200) = -1$. Two separate plots are convenient, one for the region for $|s| < 10$ and the other for $|s| > 10$. For $|s| < 10$ what single pole is approximately equivalent to the poles at -40, -100, and -200? For $|s| > 10$, what single pole is approximately equivalent to the pair of poles at the origin and the zero at -2. Sketch each plot.

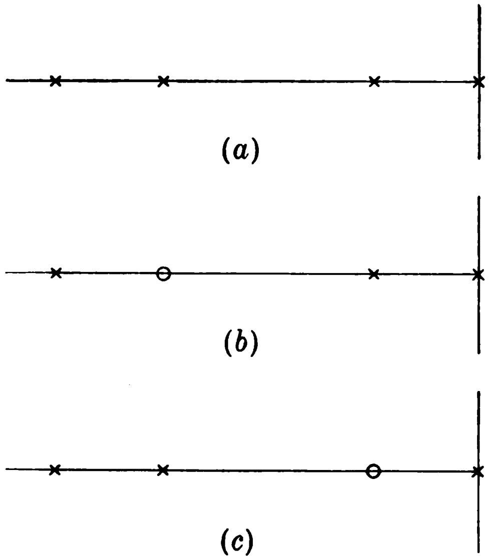


FIG. P7.1. Examples of poles and zeros along the real axis.

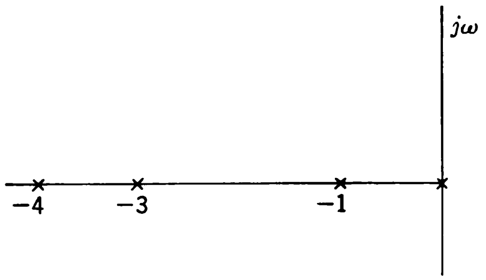


FIG. P7.2. A special problem for sketching a root-locus plot.

7.3. Sketch the locus of roots for the system shown in Fig. P7.2. Identify the angles of the asymptotes and any special points useful in establishing the locus.

7.4. The value of gain needed to make a particular point on a locus be a root can be estimated by inspection of line lengths.

(*a*) Estimate the numerical value of K_1, K_2, K_3, K_4 for $T_1 = 1$ sec and $T_2 = \frac{1}{4}$ sec in Fig. 7.8.

(*b*) Give these values of gain for $T_1 = \frac{1}{10}$ sec and $T_2 = \frac{1}{40}$ sec.

7.5. The logarithmic scale in Fig. 7.16 can be used to compute the gain needed for a given point on a root-locus plot even if unity on the plot is not the same as unity on the scale. The magnitude of a term $1 + Ts$ is the ratio of lengths $(1/T_1 + s)/(1/T_1)$, and the decibel value of $1 + T_1s$ is the difference of the decibel values of these lengths. The numerical value of s is the ratio of the length of the s vector to the length on the plot equal to 1. Use the scale shown in Fig. 7.16 to compute the gain K_2 for $T_1 = 1$, $T_2 = \frac{1}{4}$ sec for the system with poles shown in Fig. 7.8.

7.6. Prove that the locus shown in Fig. 7.12 is a circle about the point $-1/T$ as a center.

7.7. Compute the gain K_1, K_2, K_3, and K_4 needed for the roots to be at the points shown in Fig. 7.8 for $T_1 = 1$ sec and $T_2 = 4$ sec.

7.8. The transfer functions below correspond to the block diagram in Fig. 7.19. Assume that the roots of the characteristic equation are known and denote them as $r_1, r_2, r_3, \ldots$. Write the expression for the over-all transfer function in each case:

(*a*) $(\Theta_o/\Theta_a)(s) = K/s^2$; $(\Theta_f/\Theta_o)(s) = (1 + Ts)/(1 + \alpha Ts)$
(*b*) $(\Theta_o/\Theta_a)(s) = K_1/(1 + T_1s)$; $(\Theta_f/\Theta_o)(s) = K_2s$
(*c*) $(\Theta_o/\Theta_a)(s) = K_1/s^2$; $(\Theta_f/\Theta_o)(s) = K_2s$

***7.9.** The integral network in Art. 7.10 is effective in reducing the steady-state error required to cause the output to have a constant velocity. Estimate the effect of the network upon the error required for the output to have a constant acceleration.

7.10. (*a*) Estimate the value of gain K which will make the root in Fig. 7.26 be at the point r_1 for $T_1 = 1$ sec.

(*b*) Estimate the amplitude of the transient term e^{r_1t} for a unit-step input to the system.

7.11. Sketch the frequency-response curve in inverse vector form (such as Fig. 5.13) for the system whose zeros and poles are given in Fig. 7.27.

7.12. Given the system with the transfer function $K/(1 + T_1s)(1 + T_2s)s$ in which $T_1 = \frac{1}{10}$ and $T_2 = \frac{1}{40}$. For this system having a lead network $(1 + T_3s)/(1 + \alpha T_3s)$, in which $\alpha = \frac{1}{10}$, sketch the root-locus plot for the following values of T_3: (*a*) $T_3 = T_1$, (*b*) $T_3 = \frac{1}{2}T_1$, and (*c*) $T_3 = 2T_1$. (*d*) What is the gain factor for the best of the preceding plots for the root having a damping ratio of 0.4?

7.13. Given the same basic system as in Prob. 7.12. A double lead network $(1 + T_3s)^2/(1 + \alpha T_3s)^2$ is to be used in order to obtain more phase lead for a given attenuation factor, $\alpha^2 = \frac{1}{10}$.

(*a*) Sketch several plots to determine approximately the best value of T_3 for the highest possible natural-frequency root having a damping ratio of 0.4.

*(*b*) Describe a procedure to obtain the result desired in (*a*) directly rather than by trial and error.

(*c*) What is the gain factor for the best plot in (*a*) for the root having a damping ratio of 0.4?

***7.14.** Show that a lead network should be $(1 + s/\sqrt{\alpha}\omega_o)/(1 + s\sqrt{\alpha}/\omega_o)$ to obtain maximum phase lead for any root at a distance $1\omega_o$ from the origin for a given value of α.

CHAPTER 8

8.1. The block diagram in Fig. 8.1 can be solved either by closing the inner loop and then closing the outer loop as in Fig. 8.2 or by combining the two feedback signals as indicated in Eq. (8.2). Solve the system in Fig. 8.1 for $(\Theta_o/\Theta_i)(s)$ in both ways for the following values: $T_2 = 1$ sec, $T_1 = \frac{1}{4}$ sec, $K = 3$/sec, and $B = 3$ sec.

8.2. Article 8.2 shows the procedure to obtain a root-locus plot versus the amplifier time constant T_1. Solve the same system with the final root-locus plot having the motor time constant T_2 as the variable. The fixed values are $T_1 = \frac{1}{4}$ sec and $K = 2$/sec.

8.3. Solve for the roots r_1 and r_2 of the system in Fig. 7.11 for $T_1 = 1$ sec and $K = 1$/sec. Plot the locus using the values determined above versus the increase in gain ΔK. Solve for the derivative $\Delta s/\Delta K$ at $s = r_1$.

8.4. The constant sum of roots applied to the plot in Fig. 7.8 for K_3 requires that the shift of the real root, $r_3 + 1/T_2$, must be equal in magnitude to the net shift of the roots r_1 and r_2, which is $1/T_1$. Confirm this by showing that the value of gain is consistent for this entire set of roots.

8.5. The characteristic equation of a damped spring-mass system is

$$K + Ds + Js^2 = 0$$

in which K is the spring constant, D is the damping constant, and J is the inertia. Sketch the root-locus plot with (*a*) K variable, (*b*) D variable, and (*c*) J variable. In each plot take the numerical value of the other parameters to be 1. As a check, note that the position of the roots in each plot for all parameters unity must be the same.

8.6. Add a dashpot with a damping constant D_1 across the spring K_1 in Fig. 8.9. Modify the block diagrams in Figs. 8.10 and 8.11 to include this effect.

***8.7.** A combined electrical and mechanical system is involved in a voice coil driving a loud speaker or in a solenoid driving a hydraulic valve. In each case the voltage equation is of the form $V = iR + L\,di/dt + Blv$, in which v is the velocity of the movable element. The force on the element is Bli, and it causes a motion limited by mass M, damping D, and spring constant K, so that

$$f = Bli = M\frac{d^2x}{dt^2} + D\frac{dx}{dt} + Kx$$

(*a*) Set up the block diagram for such a system, showing each individual parameter separately.

(*b*) Modify the block diagram as necessary to obtain a single loop involving the feedback of the voltage Blv.

8.8. Consider a system similar to that in Fig. 8.13, except that the resonant component has some damping; the pair of poles in Fig. 8.15 are then to the left of the $j\omega$ axis.

(*a*) Sketch root-locus plots for the system for the poles having damping ratios of $\zeta = 0.1$, 0.3, 0.5, and 0.7 and for $G(s) = 1$ in the servo function.

*(*b*) What damping ratio must these poles have in order that the use of lag circuits will not increase the allowable loop gain for a stable system?

8.9. An alternate method of stabilizing the roll autopilot system is to use rate gyro feedback. Specifically another feedback loop is closed in Fig. 8.21 from θ_o through the function Bs to θ_i. The plot in Fig. 8.22 for $G(s) = 1$ determines the position of the roots before rate feedback. Raise the value of gain in this plot until one pair of roots is along the $j\omega$ axis. Make a separate plot to show the effect of rate feedback.

8.10. Alternate methods of solution of the system in Fig. 8.29 involve making separate plots of attitude feedback $K_\gamma G_\gamma$ and altitude feedback $K_y G_y$. Assume that the servo function is simply 1 in all plots.

(*a*) Sketch the root-locus plot versus K_γ with $G_\gamma = s$ and $G_y = 0$, and select K_γ such that the roots are critically damped.

(*b*) Starting from the roots determined in (*a*), sketch the locus versus K_y with $G_y = (1 + Ts)/(1 + \alpha Ts)$, in which $T = 2MV/C_L$ and $\alpha = 0.1$.

***8.11.** Find the relationship between transfer functions in Fig. 8.32 such that the numerator of the transfer function from θ_x to θ_3 is zero. Explain the method by which this result is obtained.

CHAPTER 9

9.1. The circuit in Fig. 9.1*c* is set up in block-diagram form in Fig. 9.2, and the root locus versus $1/R_1C_2$ is shown in Fig. 9.3. The problem is to study the effect of the output resistance of the source as shown in Fig. P9.1.

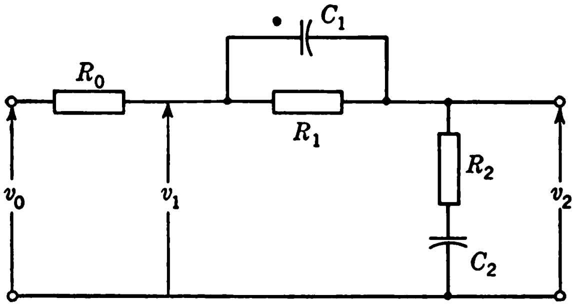


FIG. P9.1. A standard lag-lead network with impedance of the source considered.

(*a*) Modify Fig. 9.2 to include the effect of R_0.

(*b*) Sketch the root-locus plot with R_0 variable.

9.2. Assume that the source for the circuit in Fig. 9.1*b* has a resistance R_0 as in Fig. P9.1. The over-all problem is to determine the transfer function from v_0 to v_2 by inspection.

(*a*) What is the zero of the function and why?

(*b*) What is the pole of the function and why?

(*c*) Write the transfer function from v_0 to v_2.

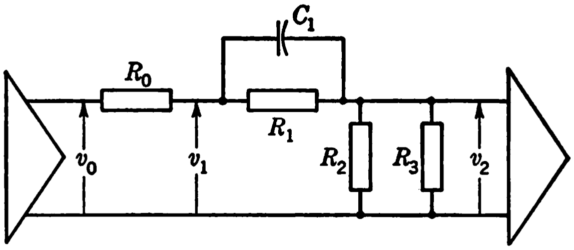


FIG. P9.2. A standard lead network with impedance of the source and the load considered.

9.3. (*a*) Find the impedance of the circuit in Fig. 9.4 as seen from the terminals of R_3.

(*b*) Sketch the root-locus plot with R_3 variable.

***9.4.** Show that, if any resistance or any capacitance in an *RC* network is increased, the time constants of all natural transients of the circuit increase.

9.5. The system in Fig. 9.17 can be studied for the effect of an output impedance of the amplifier driving the network or the input impedance of the amplifier being driven (Fig. P9.2). Note that the d-c value of impedance might be a capacitance or inductance as well as a resistance.

(*a*) What is the impedance Z_0 of the system as seen from the terminals of R_0?

(*b*) Sketch the root locus versus R_0.

(*c*) What is the impedance Z_3 of the system as seen from the terminals of R_3?

(*d*) Sketch the root locus versus R_3.

9.6. The root-locus plot in Fig. 8.17 is plotted versus loop gain for the system shown in Fig. P9.3. The purpose is to determine the optimum value of the time constant $T = RC$.

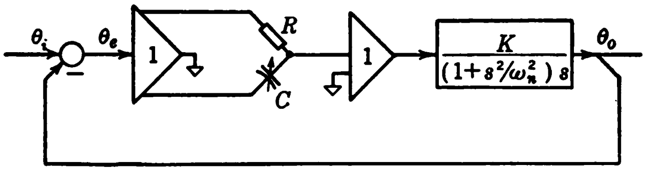


FIG. P9.3. Block diagram of a system in which the effect of capacitance C is to be determined.

(*a*) Plot the zeros and poles of the impedance of the system as seen from the terminals of the capacitor for the value $K = 0.2\omega_n$.

(*b*) Plot the locus of roots versus C.

(*c*) Locate the position of the roots on the locus for $1/RC = \omega_n/2$, the value used in the variable-gain plot (Fig. 8.17).

9.7. Construct a plot similar to that in Fig. 9.22 for the system shown in Fig. P9.4 with $T_1 = 1$ sec, $K = 1/\text{sec}$.

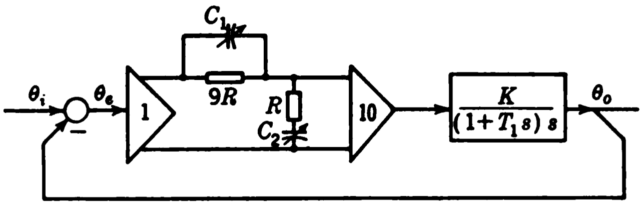


Fig. P9.4. Block diagram of system in which the effects of two capacitors are to be determined.

***9.8.** The servo system in Fig. 9.23 can be studied for another spring effect, one directly connected between the load and a fixed reference. Assume that the shaft is perfectly rigid.

(*a*) Find the mechanical impedance of the system as seen from the spring.

(*b*) Plot the locus of roots versus the load spring constant K.

(*c*) If the load is an aileron, the spring constant K may be negative under certain conditions; therefore sketch the locus for negative values of K.

9.9. (*a*) Draw a complete block diagram for the system in Fig. 9.17 in which the capacitor C_1 appears as an isolated parameter in one path.

(*b*) Write a set of equations for the system in Fig. 9.17 such that the parameter C_1 appears only once in the entire set of equations.

CHAPTER 10

10.1. Any of the systems previously studied can be set up in the form shown in Fig. 10.2. Set up each of the following systems with the order of the variables in the sequence given:

(*a*) Figure 9.11: v_1, i_1, i_2, v_2.

(*b*) Figure 9.11: Add to the array developed in (*a*) above the elements necessary to describe the effect of current i_3 which can flow through a load resistance R_3 across v_2 around the loop which includes C_1 and R_2.

(*c*) Figure 9.25: θ_i, θ_M, θ_L, m_L.

10.2. Reduce the system given in Fig. 10.1 to a form in which the variable in the last column does not affect the system (*a*) by manipulation of rows only; (*b*) by manipulation of columns only. In each case indicate the significance of the result.

10.3. An array of coefficients can be considered to be a block diagram, as shown in Fig. 10.2. A particular transfer function around a closed loop is shown in Fig. 10.7.

(*a*) Write all the loop transfer functions for the array in Fig. 10.7 for the special case of a_{31} and a_{32} equal to zero.

(*b*) Write the additional functions introduced by a_{31}.

(*c*) Write the additional terms introduced by a_{32}.

(*d*) Give a minus sign to each term which involves the principal diagonal an odd number of times. Add -1 to the sum of the terms in (*a*), (*b*), and (*c*), and show that the result is the same as that obtained by expansion of a determinant by any method.

10.4. (*a*) Solve for the transfer function from Λ_o to Γ for the array of coefficients in Fig. 10.1 by use of Cramer's rule.

(*b*) Solve for the same transfer function by treatment of the system in block-diagram form.

10.5. The pattern of solution using transfer functions, as indicated in Fig. 10.12, can be set up for more complicated systems by using equations of the form of (10.23) or (10.24).

(*a*) Set up the equations to determine the roots for the system in Fig. 10.12 including the coefficient C_{51}.

(*b*) Set up the equations of the system in Fig. 10.12 with $C_{51} = 0$ but C_{42} present.

10.6. Given the system defined by the set of equations in Fig. P10.1, in which C identifies fixed functions and F identifies adjustable functions.

$$\begin{array}{cccc} & \Theta_1 & \Theta_2 & \Theta_3 \\ X_1 & C_{11} & F_{12} & F_{13} \\ X_2 & -1 & C_{22} & F_{23} \\ X_3 & -1 & F_{32} & C_{33} \end{array}$$

FIG. P10.1. Determinant with given function denoted by C and adjustable functions by F.

(*a*) What conditions must be satisfied in order that the input X_1 may not produce any output Θ_3?

(*b*) What conditions need be satisfied in order that neither X_1 nor X_2 may produce any output Θ_3?

(*c*) What is the transfer function from X_3 to Θ_3 if conditions in (*b*) are satisfied?

10.7. Given the system shown in Fig. 10.12.

(*a*) What is the addition to the determinant due to the addition of the coefficient C_{15}?

(*b*) What is the addition to the determinant if the coefficient C_{51} is added with the product $C_{15}C_{51}$ neglected?

(*c*) What is the addition to the determinant involving the product $C_{51}C_{15}$?

***10.8.** Given a very complicated system whose characteristic equation is known in the form $F_1(s) + CF_2(s) = 0$, where $F_1(s)$ and $F_2(s)$ are known polynomials in s. Describe the procedure by which a small change in the position of the known root ($s = j\omega_1$) can be determined for a small change in C.

CHAPTER 11

11.1. Given the system whose open-loop transfer function is

$$\frac{\Theta_o}{\Theta_e}(s) = \frac{K(1+s)}{s(1+s/2)(1+s/10)^2}$$

Find the amplitude A_1 of the transient term $A_1e^{r_1t}$, in which r_1 is the root between the origin and -1, for the following values of gain K: (*a*) 0.1; (*b*) 0.5; (*c*) 2.5.

11.2. Given the quadratic system shown in Fig. 11.5. Plot the transient response of the system for a unit-step input for the following values of roots:

(*a*) $r_1 = r_2 = -0.5$.

(*b*) $r_1 = -0.6$, $r_2 = -0.4$. Use the general formula for nonrepeated roots.

(*c*) $r_1 = -0.6$, $r_2 = -0.4$. Use a series type of formula, such as (11.17).

11.3. The vector plot shown in Fig. 11.7 can be modified by first fixing the projection axis for amplitude as the positive real axis and then locating the vector whose projection represents the transient.

(*a*) Show the locus of points for the head of the vector such that the initial amplitude is A.

(*b*) Locate the position of the vector such that the initial amplitude of the oscillation is A and the initial velocity is B.

11.4. Plot the transient response given in (11.8).

***11.5.** Approximate the transfer function needed to create the transient in Fig. P11.1 for a unit-step input.

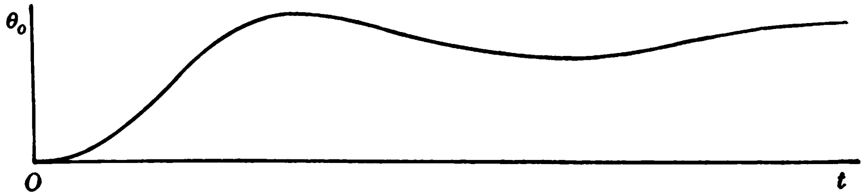


Fig. P11.1. A transient produced by a unit step on a linear system.

11.6. Sketch the transient response for a unit-step input using the method of Art. 11.7 for the system given in Prob. 11.2. (*a*) $1/(1 + s/0.5)^2$. (*b*) $1/(1 + s/0.4)(1 + s/0.6)$.

11.7. Given the arbitrary input shown in Fig. P11.2 for a system whose transfer function is $1/(1 + s^2/\omega_n^2)$.

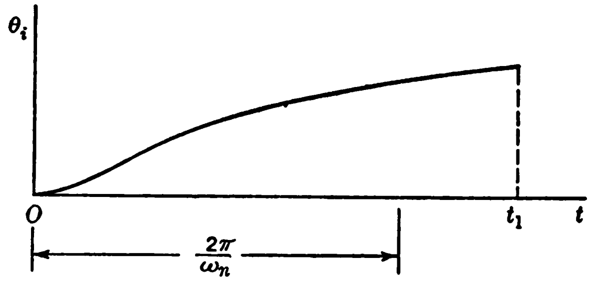


Fig. P11.2. An arbitrary function of time.

(*a*) Draw the output of the system using approximately five step inputs to approximate the curve up to $t = t_1$.

(*b*) Draw the output using ramps to approximate the curve as closely as the step inputs in (*a*).

11.8. Using graphical convolution, solve for the output at $t = t_1$ in Fig. P11.2 for a system transfer function $1/(s^2 + \omega_n^2)$.

***11.9.** Solve the problem stated in Prob. 11.7 using numerical convolution with the following approximations of the sine-wave weighting function: (*a*) 0, 1, 0, −1, 0, 1 . . . ; (*b*) 0, 1, 1, 0, −1, −1, 0, 1 . . . .

CHAPTER 12

12.1. Sketch the remainder of Fig. 12.3 until an equilibrium condition is reached.

12.2. Phase-plane curves can be sketched by knowing the slope of the curve, $dv/d\theta$, as a function of v and θ. The time variable must, of course, be eliminated by a substitution, such as $v = d\theta/dt$. In each case, solve for $dv/d\theta$ and then sketch the phase-plane curve.

(*a*) $(1 + s^2)\theta = 0$.

(*b*) $(1 + 2\zeta s + s^2)\theta = 0$; $\zeta = 0.5$.

12.3. Sketch the family of curves which show the period of oscillation of the system in Fig. 12.8 as a function of the ratio F/K and the amplitude θ_o.

12.4. Sketch the velocity-displacement curves for the system in Fig. 12.16 with the function $K/(1 + Ts)s$ replaced by K/s^2.

12.5. Sketch the attenuation characteristic versus amplitude for a nonlinear function of the form (*a*) $\theta_o = \theta_i^2$; (*b*) $\theta_o = \theta_i^{\frac{1}{2}}$.

BIBLIOGRAPHY

The following references are listed according to the chapter in which the material is most applicable.

Chapter 1

Bollay, Wm. E.: Aerodynamic Stability and Automatic Control, *J. Aeronaut. Sci.*, September, 1951.

Hazen, H. L.: Theory of Servomechanisms, *J. Franklin Inst.*, Vol. 218, pp. 279–331, September, 1934.

Moore, J. R.: Application of Servo Systems to Aircraft, *Aeronaut. Eng. Rev.*, Vol. 8, pp. 32–43, January, 1949.

Chapter 2

Greenwood, Ivan A., Jr., J. Vance Holdam, Jr., and Duncan MacRae, Jr., eds.: "Electronic Instruments," Vol. 21, MIT Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1948.

Chapter 3

Johnson, Walter C.: "Mathematical and Physical Principles of Engineering Analysis," McGraw-Hill Book Company, Inc., New York, 1944.

Lauer, Henri, Robert Lesnick, and Leslie E. Matson: "Servomechanism Fundamentals," McGraw-Hill Book Company, Inc., New York, 1947.

Skilling, Hugh H.: "Transient Electric Currents," 2d ed., McGraw-Hill Book Company, Inc., New York, 1952.

Chapter 4

Gardner, Murray F., and John L. Barnes: "Transients in Linear Systems," John Wiley & Sons, Inc., New York, 1942.

Churchill, Ruel V.: "Modern Operational Mathematics in Engineering," McGraw-Hill Book Company, Inc., New York, 1944.

Chapter 5

Ahrendt, W. R., and John F. Taplin: "Automatic Feedback Control," McGraw-Hill Book Company, Inc., New York, 1951.

Bode, H. W.: "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, Inc., New York, 1945.

Brown, G. S., and D. P. Campbell: "Principles of Servomechanisms," John Wiley & Sons, Inc., New York, 1948.

Chestnut, Harold, and Robert W. Mayer: "Servomechanisms and Regulating System Design," John Wiley & Sons, Inc., New York, 1951.

James, Hubert M., Nathaniel B. Nichols, and Ralph S. Phillips: "Theory of Servomechanisms," McGraw-Hill Book Company, Inc., New York, 1947.

Marcy, H. Tyler: "Parallel Circuits in Servomechanisms," *Trans. AIEE*, Vol. 65, pp. 539–545, 1946.

Smith, Ed S.: "Automatic Control Engineering," McGraw-Hill Book Company, Inc., New York, 1944.

Chapter 6

Chestnut, Harold, and Robert W. Mayer: "Comparison of Steady-state and Transient Performance of Servomechanisms," *Trans. AIEE*, Vol. 68, pp. 765–777, 1949.

Evans, Walter R.: "Graphical Analysis of Control Systems," *Trans. AIEE*, Vol. 67, pp. 547–551, 1948.

Harris, H., Jr., M. J. Kirby, and E. F. von Arx: "Servomechanism Transient Performance from Decibel-Log Frequency Plots," *AIEE Tech. Paper* 51–269, 1951.

Profos, P.: "A New Method of Regulating System Design," *Sulzer Tech. Rev.* (*Switz.*), No. 2, 1945.

Chapter 7

Chu, Yaohan: Feedback Control Systems with Dead-time Lag or Distributed Lag by the Root Locus Method, *Trans. AIEE*, Vol. 71, Part II, pp. 291–296, 1952.

Evans, Walter R.: "Control System Synthesis by the Root Locus Method," *Trans. AIEE*, Vol. 69, pp. 66–69, 1950.

Chapter 8

Aaron, M. R.: "Synthesis of Feedback Control Systems by Means of Pole and Zero Location of the Closed Loop Function," *AIEE Tech. Paper* 51–267, 1951.

Moore, J. R.: "Combination Open-Cycle Closed-Cycle Systems," *Proc. IRE*, Vol. 39, No. 11, pp. 1421–1432, 1951.

Chapter 9

Evans, Walter R.: "Application of Root Locus Method to Multicoupled Systems," thesis, University of California, 1951.

Chapter 10

Chu, Yaohan: Correlation between Frequency and Transient Response of Feedback Control Systems, *Trans. AIEE*, Vol. 72, Part II, pp. 81–92, 1953.

Doherty, Robert E., and Ernest G. Keller: "Mathematics of Modern Engineering," John Wiley & Sons, Inc., New York, 1944.

Chapter 11

Bubb, Frank W.: "Cardinal Spectrum Analysis of Physical Systems," Office of Air Research Report, Wright Field, Ohio, 1951.

Chapter 12

Jackson, Kenneth R.: "Investigation of the Effect of Error plus Error-rate Equalization of Shunt Motor Relay Servomechanisms," thesis, University of California at Los Angeles, 1951.

Kochenburger, R. J.: "A Frequency Response Method of Analyzing and Synthesizing Contactor Servomechanisms," *Trans. AIEE*, Vol. 69, pp. 270–283, 1950.

McCann, G. D., F. C. Lindvall, and C. H. Wilts: "The Effect of Coulomb Friction on the Performance of Servomechanisms," *Trans. AIEE*, Vol. 67, 1948.

Rogers, T. A., and W. C. Hurty: "Relay Servomechanisms, the Shunt Motor Servo with an Inertia Load," *Trans. ASME*, Vol. 72, pp. 1163–1172, 1950.

Weiss, H. K.: "Analysis of Relay Servomechanisms," *J. Aeronaut. Sci.*, Vol. 13, No. 7, pp. 364–376, July, 1946.

Appendix A

Gardner, Murray F., and John L. Barnes: "Transients in Linear Systems," John Wiley & Sons, Inc., New York, 1942.

Appendix B

Chu, Yaohan: Synthesis of Feedback Control Systems by Phase Angle Loci, *Trans. AIEE*, Vol. 71, Part II, pp. 330–339, 1952.

Yeh, Victor C. M.: The Study of Transients in Linear Feedback Systems by Conformal Mapping and the Root Locus Method, *Trans. ASME*, Vol. 76, No. 3, pp 349–361, 1954.

Appendix D

Guillemin, Ernest A.: "Mathematics of Circuit Analysis," John Wiley & Sons, Inc., New York, 1949.

Routh, E. J.: "Dynamics of a System of Rigid Bodies," 3d ed., Macmillan & Co., Ltd., London, 1877.

Appendix E

Bewley, L. V.: "Two-dimensional Fields in Electrical Engineering," The Macmillan Company, New York, 1948.

Appendix G

James, Hubert M., Nathaniel B. Nichols, and Ralph S. Phillips: "Theory of Servomechanisms" McGraw-Hill Book Company, Inc., New York, 1947.

INDEX

Aaron, M. R., 275
Accuracy, pick-offs, 16
  standard inputs, 7
Ahrendt, W. R., 275
Amplifier, a-c, 17–19
  d-c, 17–19
Amplitude of transient terms, 45
  from conformal map in $F(s)$ plane, 90
  from initial conditions, 192
  for nearly equal roots, 190
  from root-locus plots, 188
  for unit-impulse input, 55
  from zeros and poles, 186
Arx, E. F. von, 276
Asymptote of root-locus plots, 19, 129
Attenuation diagram, 73–80,
Autocorrelation functions, 253–262
  factors of, 261
  graphical picture, 254

Back-emf constant, 21
Backlash, 215–217, 220
Barnes, J. L., 275, 277
Bewley, L. V., 277
Block diagrams, 48
  of circuits, 146
  elimination of variables, 142–145
  interchange of blocks, 54, 55
  replacement of loop, 111
Bode, H. V., 275
Bode diagram, 73–80,
Bollay, W. E., 275
Break frequency, 74
Bridge circuit, phase-shifting, 134
Bridged-T networks, 156, 157
  effect on modulated carrier, 159, 160
  locus of zeros, 157
Brown, G. S., 275
Bubb, F. W., 276

Campbell, D. P., 275
Characteristic equations, 61–63, 96–99
Chestnut, H. E., 275, 276
Chu, Y., 276, 277
Churchill, R. V., 276, 277
Closed-loop frequency response, 70–72
Closed-loop transfer function, 110–113
  zero of, 112
Compensation, active, 135
  (*See also* specific types of networks)
Complex convolution, 257, 258
Complex numbers, 41, 42
  derivative of, 248
  (*See also* Vectors)
Compliance, effect on roots, 165
Conditional stability, 69
Conformal mapping, 88–90
  derivative from $F(s)$ plot, 91
  of electric field, 247
  in the s plane, 248
Conjugate terms, 188
Control systems, types of, 2
Conversion charts, 76–78
Convolution, 199
  complex, 257, 258
  graphical, 228
  inverse, 203
  numerical, 201
Cramer's rule, 174–177
Curve fitting, 82

Damping, of motor, 31
  position of roots, 45
  (*See also* Quadratic functions)
Damping ratio, 46
Decibels, 73
Demodulators, 19–21
Derivative of complex number, 248
Derivative circuit (*see* Lead network)
Determinants, 170–185

Determinants, addition of rows and columns, 172
  compared to block diagrams, 170
  Cramer's rule, 174–177
  minors, 176
  *n*th-order, 183
  simplified expansion, 181
Dickensen, W. E., 236*n*.
Differential equations, 61, 96–99
  graphical solution, 198
  (*See also* Transfer functions)
Doherty, R. E., 276

e, numerical value of, 34
E pick-offs, 16
Electric-field analogy, 247, 248
Equalization, 135
  (*See also* specific types of networks)
Error, 1
  dynamic, 8
  mean square (*see* Mean square error)
Euler's equation, 43
Evans, W. R., 276
Exponential functions, 33, 34

Feedback functions, 59
  effect of, on frequency response, 72
    on roots, 110–113
  tachometer, 122–124
Flux plot (*see* Conformal mapping)
Foster's theorem, 155
Fourier series, 249
Frequency response, 10, 63
  closed-loop, 70–72
  effect of feedback functions on, 72
  inverse vector plot, 71
  logarithmic plots, 73–80
  nonlinear, 217–220
  open-loop, 67, 68
  transfer functions from, 82
  zeros and poles, 120

Gain, effect on roots, 102–105
  logarithmic plots, 75, 77
  small variations in, 125–127
Gardner, M. F., 275, 277
Gear ratio, choice of, 27–29
Graphical solution, of differential equation, 198
  of polynomial, 183, 184
  (*See also* Conformal mapping; Frequency response; Root-locus plots)
Greenwood, I. A., Jr., 275
Guillemin, E. A., 277

Harris, H., Jr., 276
Hazen, H. L., 275
Holdam, V., Jr., 275
Hunting (*see* Stability)
Hurty, V. C., 276
Hydraulic components, 26

Impedance, 146–148
  of complete systems, 161
  of LC networks, 155
  of RC networks, 153–155
Induction-motor characteristic, 23
Initial conditions, 57, 192
Integral network, 78
  effect of, on logarithmic plot, 79
    on root-locus plot, 115
    on transient, 117
  effect on, of resistance load, 150
Integration in the complex plane, 228
Inverse Laplace transform, 226
Inverse transfer function, 71
Inverse vector plot, 71

Jackson, K. R., 277
James, H. M., 275, 277
Johnson, W. C., 275

Keller, E. G., 276
Kirby, M. J., 276
Kochenburger, R. J., 276

Lag-lead network, 147, 148
  effect of load on, 151–153
Laplace transform, 225–229
  complex convolution, 257
  inverse, 226
  physical concept of, 229
Lauer, H., 275

Lead network, 79
  effect of, on logarithmic plot, 80
    on root-locus plot, 105, 119
  90° locus in s plane, 118
Lesnick, R., 275
Limit cycle, 213
Lindvall, R. C., 276
Locked controls, 140, 141
Locus (*see* Frequency response; Root-locus plots)
Logarithm of a complex function, 248
Logarithmic plots, 73–80
  effect of lead network on, 80
Logarithmic scale of vector lengths, 110

M circles, 71
McCann, G. D., 276
Maclaurin series for e, 34
MacRae, D., Jr., 275
Magnification factor, 71
Mapping (*see* Conformal mapping)
Marcy, H. T., 275
Matson, L. E., 275
Mayer, R. W., 275, 276
Mean square error, 249, 250
  as double convolution integral, 253
  minimum value of, 259
  transient build-up of, 256
Mechanical analogies, 30–32
Minors, 76
Modulated carrier signal, 158–160
Modulators, 19–21
Moore, J. R., 275, 276
Motor, a-c, 22
  d-c, 21
Multicoupling, 142
  (*See also* Determinants)
Multiple-loop systems, 122–142

Nepers, 94
Networks (*see* specific types of networks)
Nichols, N. B., 275, 277
Nichols chart, 76–78
Noise, frequency components of, 249–253
  (*See also* Autocorrelation functions)
Notch network (*see* Bridged-T networks)
Nyquist's criterion of stability, 69, 85–88

Open cycle-closed cycle system, 8
Open-loop response, 67, 68
Over-all transfer function, 110–113

Phase-plane diagrams, 207–216
Phase-shifting bridge circuit, 134
Phillips, R. S., 275, 277
Pick-offs, 15–17
Pitch autopilot, 139–145, 178, 179
Pneumatic components, 27
Pole, 103
  by inspection, 148–150

Quadratic functions, frequency-response, 83
  frequency-response curves, 65
  transient-response curve, 46

RC networks (*see* Integral network; Lag-lead network; Lead network)
Regulator, 2
Relay time delay, 212, 213
Resonant system, inner-loop, 133
  phase-plane diagram, 208–210
  spring-mass, 36–38
RLC network, 135, 136
Roger, T. A., 276
Roll-yaw coupling, 180–182
Root-locus plots, 100–108
  aids in sketching, 230–238
  amplitude of transient terms from, 188
  asymptote of, 19, 129
  effect on, integral network, 115
    of lead network, 105, 119
  use of spirule, 239–242
Roots, classical method, 96–99
    characteristic equation, 61–63
  effect of, on transients, 104
  effect on, of compliance, 165
    of feedback functions, 110–113
    of gain, 102–105
  graphically on logarithmic plots, 183–185
  position in damping, 45
  vector sum constant, 127, 128
  (*See also* Root-locus plots)
Routh, E. J., 277
Routh's criterion, 63, 242–246

Servomechanisms, 2
Sinusoidal response (*see* Frequency response)
Skilling, H. H., 275
Small variation in gain, 125–127
Smith, E. S., 276
Solenoid, 24
Spirule, 110, 237–242
  use of, in root-locus plots, 239–242
Spring-mass system, 30–32, 36–38, 43–45
Stability, 1
  conditional, 69
  criterion of, Nyquist's, 69, 85–88
    Routh's, 63, 242–247
Steady-state analysis (*see* Frequency response)
Superposition, of feedback signals, 59, 60
  simplification allowed, 9, 10
  in Thévenin's theorem, 150, 151
  in weighing function, 199–201

Tachometer feedback, 122–124
Taplin, J. F., 275
Thévenin's theorem, 150, 151
  for choice, of capacitance, 160
    of compliance, 165
Time constant, 32, 33
Torque-speed characteristics, 22, 23, 31
Transfer functions, 50
  closed-loop, 110–113
  from frequency response, 82
  by inspection, 140–150
  inverse, 71
  from transient response, 195
Transform (*see* Laplace transform)
Transient response, 30
  from frequency response, 82
  due to impulse input, 52
  sketching aids for, 193–195
  transfer functions from, 195
  (*See also* Amplitude of transient terms)

Undamped natural frequency, ω_0, 67
Unit impulse, 52
Unit step, 53

Vectors, and complex numbers, 41, 42
  derivative of, 39
  plot of open-loop function, 99
  for sine waves, 38
  sum of roots, 127, 128
Vibration absorber, 130–132
von Arx, E. F., 276

Weighing function, 199
Weiss, H. K., 276
Wilts, C. H., 276

Yeh, V. C. M., 277

Zero, 105
  of closed-loop function, 112
  by inspection, 148–150

67 226 A A 30

NIVERSITY

brought to the d

UNIVERSITY OF MICHIGAN
3 9015 00293 3185